Architectural Analysis of a Baseline ISP Pipeline

Hyun Sang Park

Abstract An ISP is an entity that performs various image-processing algorithms on a raw image from an image sensor. A number of functions are incorporated in an ISP, and they are combined together similarly but differently among ISP implementers. ISP functions are divided into pixel-based and frame-based ones, and are dedicated to one of three color domains in Bayer, RGB, or YCbCr. Although it is an essential component for a camera system, surprisingly, its architecture has not been analyzed in the context of standards. The purpose of this chapter is to remove ambiguity when analyzing an ISP architecture or designing a new ISP architecture. At the end of this chapter, a baseline ISP pipeline is presented, which is tentatively built to conform to the existing standards.

Keywords Image signal processor · Image pipeline · Image sensor · Bayer sensor

1 Introduction

The functions implemented in ISP can be categorized into two groups. The first includes pixel-based functions. It makes the result by utilizing an input pixel and its surrounding pixels. It is also regarded as a spatial filter because its output is generated by exploiting spatial information. The second contains frame-based functions. To obtain the processed result, these functions require the whole pixels of an image. Frame-based functions are further divided by how many images are exploited to get the outcome.

One is to refer to global features of the whole frame of a single image. The image quality of an image needs to be consistent over all portions of the image. The method for extending the dynamic range of an image can be included in this category. There are many other algorithms such as auto-white balance,

H.S. Park (✉)
Division of Electrical Electronics and Control Engineering, Kongju National University, Gongju, South Korea
e-mail: vandammm@kongju.ac.kr

C. Kyung (ed.), *Theory and Applications of Smart Cameras*,
KAIST Research Series, DOI 10.1007/978-94-017-9987-4_2

auto-exposure, contrast enhancement, which extract the global features from the given single image. The other functions that require a plural number of image frames often utilize temporal correlation among them. Some algorithms to reduce noise or distortion are included in this group. They analyze the temporal correlation between frames, and include following algorithms such as temporal noise reduction, rolling-shutter removal, image stabilization, and so on.

Frame-based functions are not handled in traditional ISPs except for auto-exposure control, auto-white balance, and auto focus (also known as 3A or 3-auto). For example, if noise is to be reduced by considering temporal correlation, at least two image frames have to be stored to check if it can be regarded as noise or not. Basically, an ISP has been developed to be embedded in an image sensor. Because of this requirement, it cannot work with functions requiring the frame memory. The 3A algorithm doesn't need the frame memory because the global features that 3A requires can be extracted while scanning the current frame. Although they are regarded as frame-based ones, they could be considered as basic components in the traditional ISP architecture since they do not need the frame-memory itself. In general an ISP can be implemented in three ways.

1.1 Embedded ISP in an Image Sensor

It is what is called the baseline ISP, which has a cascaded pipeline architecture composed of spatial filters and point functions. Allowed frame-based functions are limited to only 3A algorithms, which do not require any frame-memory.

1.2 Discrete ISP Package

In the early era of ISP commercialization, a baseline ISP itself was built solely as a discrete chip. These days it is often produced in a multi-chip package with a stacked SDRAM as frame memory. Because it embeds the frame memory inside, it can support frame-based functions such as image stabilization, temporal noise reduction, wide dynamic range, and so on. However, it still has difficulties in handling those algorithms derived from computer vision technology, which also utilize the images stored in the frame memory but require large number of floating-point operations and complicated control flow. Adopting power-consuming CPU and/or GPGPU is not considered yet.

1.3 Embedded ISP Inside an AP

There are powerful programing units like CPU/GPGPU inside an application processor (AP). Besides, the application processor provides abundant memory

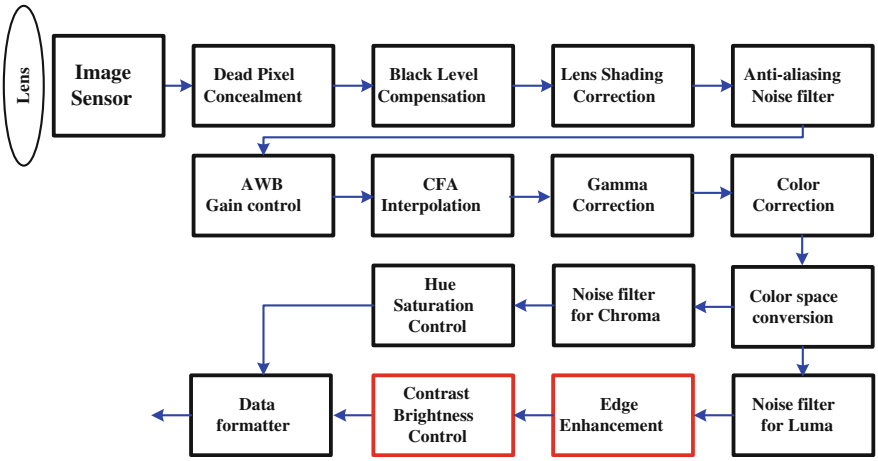


Fig. 1 The definitive form of a baseline ISP pipeline

space as well as bandwidth. So the pixel-based functions can be processed with a legacy baseline ISP, while the frame-based functions can be processed by programing GPU/GPGPU. This form of the ISP implementation consumes much energy since it uses the power-hungry memory device and the hot computing units. Nevertheless, it can provide the best quality of an image for end-user satisfaction.

The pipelined chain of an ISP is not standardized, such that each implementer has devised lots of very similar, yet different ISP pipelines. In this section, the baseline ISP in Fig. 1 will be discussed in the context of known standards.

2 Primary ISP Architecture for Bayer Image Sensors

The ISP itself is not a subject under standardization, but the standardization of digital video has been built continuously for a long time. Rec. ITU-R Rec. 601 [1] and Rec. ITU-R BT. 656 [2] (also known as CCIR601/656) constituted in 1982 claims the standardization of basic component of an ISP for the first time.

The camera module in Fig. 2 consists of an optical module, an image sensor, and an ISP. The ISP here contains three components: quantization, color space

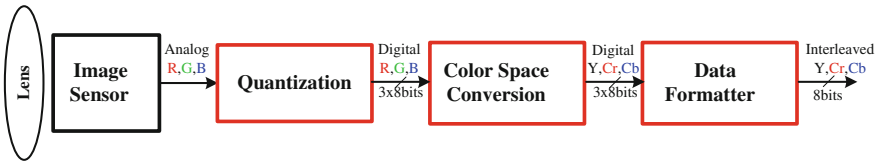


Fig. 2 Camera module architecture in Rec. ITU-R BT.601 and Rec. ITU-R BT.656

conversion, and data formatter. The image sensor is assumed to produce analog R, G, and B signals at every pixel position. In Rec. ITU-R BT.601, the first two functions are standardized, and in Rec. ITU-R BT.656 the last function is standardized.

The title of Rec. ITU-R BT.601 is "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios" and defines common regulations on digitization of digital video for SDTV (Standard Definition Television). Video in this standard has the resolution of 720 × 480 or 720 × 576 at the sampling frequency of 13.5 MHz. This recommendation standardizes how to obtain the corresponding digital video data. When analog R, G, and B signals—E_R, E_G, and R_B—of the 1.0 volt dynamic range are given, 8-bit digital RGB signals are quantized as below. They will have 219 values which reside between 16 and 235.

$$\begin{aligned} E_{R_D} &= \text{int}(219E_R) + 16 \\ E_{G_D} &= \text{int}(219E_G) + 16 \\ E_{B_D} &= \text{int}(219E_B) + 16 \end{aligned} \tag{1}$$

Y, C_R, and C_B signals are calculated from these digital R, G, and B signals. The formula to convert R-G-B into Y-C_R-C_B is defined a little bit differently according to the recommendations. For example, Rec. ITU-R BT.709 [3] and Rec. ITU-R BT.2020 [4] specify the digital video format for HDTV (High Definition Television) and UDTV (Ultra Definition Television) in a very similar way to Rec. ITU-R BT.601. Although these recommendations standardize the digital video formats at difference resolutions, their color space conversion to the Y-C_B-C_R space is not the same. That is, there is no color compatibility between them. In case of inverse transformation from Y-C_R-C_B made by a different regulation to R-G-B, there may be some differences among reconstructed R-G-B data. So any ISP implementer should obey the formula specified in the appropriate recommendation. Equation (2) is what is recommended in Rec. ITU-R BT. 601. Each arithmetic operation in Eq. (2) is designed to be implemented by integer operations. Allowing the use of integer operations gives consistent calculation results among different implementations in hardware or software.

$$\begin{aligned} Y &= \frac{77}{256}E_{R_D} + \frac{150}{256}E_{G_D} + \frac{29}{256}E_{B_D} \\ C_R &= \frac{131}{256}E_{R_D} - \frac{110}{256}E_{G_D} - \frac{21}{256}E_{B_D} + 128 \\ C_B &= -\frac{44}{256}E_{R_D} - \frac{87}{256}E_{G_D} + \frac{131}{256}E_{B_D} + 128 \end{aligned} \tag{2}$$

In Rec. ITU-R BT.601, subsampling is performed horizontally with C_B and C_R components after color conversion. There exist some subsampling formats on Y-C_R-C_B signals, such as 4:4:4, 4:2:2, 4:1:1, or 4:2:0. These subsampling formats are available only in the Y-C_R-C_B color space, not for legacy RGB color spaces. The

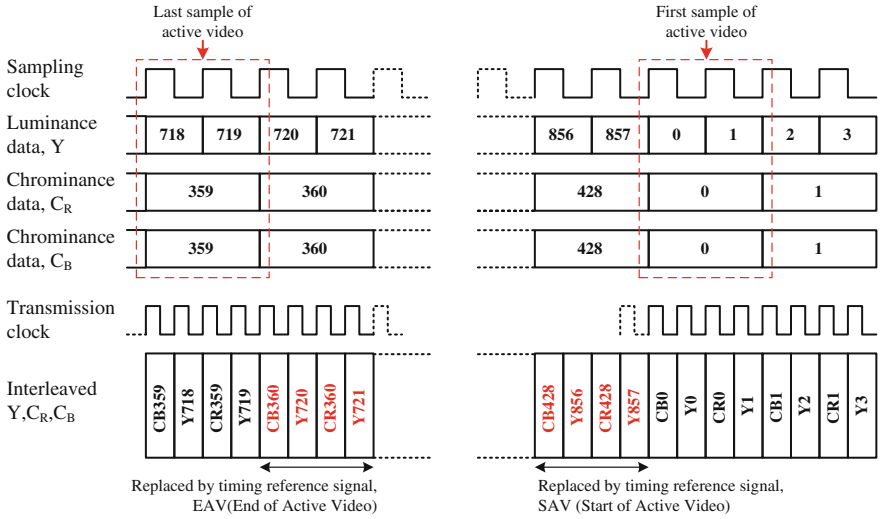


Fig. 3 Interleaved Y-C_R-C_B data format by Rec. ITU-R BT.656

subsampling on C_B-C_R components is desirable when effective data reduction is required without loss of visual quality degradation. There are large correlations between R, G, and B signals, but a few between C_R and C_B signals. Rec. ITU-R BT.601 only regulates 4:4:4 and 4:2:2 chroma subsampling formats. The 4:4:4 chroma subsampling format represents that there is indeed no subsampling. The 4:2:2 chroma subsampling format allows the subsampling on C_B and C_R chroma signals with 2:1 horizontally. The corresponding subsampled Y-C_R-C_B data are then interleaved into a single data stream according to Rec. ITU-R BT.656. The module to do subsampling and to interleave Y-C_R-C_B signals is Data Formatter in Fig. 2. The formatted data by Data Formatter will have the form like that in Fig. 3. In Rec. ITU-R BT.656, only the chroma 4:2:2 subsampling format is allowed. Thus the standardized digital video produced by conventional camera modules always supports the chroma 4:2:2 subsampling format.

Data formatter of an ISP needs the output speed to be twice as fast as any other part in the ISP, instead of using a number of data signals. An ISP usually has two clock domains. For example, a camera module made by Rec. ITU-R BT.601 and Rec. ITU-R BT.656 takes 13.5 MHz as the sampling frequency and 27.0 MHz as the output data frequency, respectively. The transferred signals through Rec. ITU-R BT.656 are only video, and no timing reference signals that define horizontal/vertical blanking periods are explicitly transferred. Instead those timing reference signals are derived from video data, where some reserved codewords are inserted at appropriate locations within the data stream.

In Fig. 3, a line of an image frame consists of 858 luminance (Y) data. Among them, the number of valid video data is 720. The interval of producing invalid data is called the horizontal blanking period. In Rec. ITU-R BT.656, four successive

words just before and after the valid data are replaced by a reserved codeword sequence such that correct timing reference signals can be derived. The four codewords substituted at the end of a valid line are called EAV (End of Active Video) and those before the beginning of a valid line are called SAV (Start of Active Video). Each codeword in SAV or EAV has either 8-bit or 10-bit, but 8-bit is preferably used in industries. SAV and EAV have the sequence of 'FF-00-00-XY' in hexadecimal numbers.

The first three codewords constitute a synchronization code to inform the receiver of the existence of timing reference. Because the synchronization code is used to synchronize the communication between a transmitter and a receiver, the synchronization code itself cannot happen by chance in video data. Otherwise erroneous synchronization will happen, which will result in a failed reconstruction of an image. The emulation of the synchronization code will not be made in practice. If Rec. ITU-R BT.601 is used in producing the digital video data, no '00' or 'FF' is allowed to be generated. In practical implementation of an ISP, however, it is often necessary to consider at the transfer stage not to emulate the synchronization code because the ISP may use all 256 values that an 8-bit code can have. There are three timing information signals such as F, V, and H, where they are transferred with protection bits at the last codeword of SAV or EAV. Their bit positions and meanings are given in Table 1.

As described above, the simplest form of an ISP is composed of quantization, color space conversion from R-G-B to Y-C_R-C_B, and data formatter. All of these steps are standardized in Rec. ITU-R BT.601 and in Rec. ITU-R BT.656, respectively. Because the quantization step is mostly embedded in an image sensor, the minimum number of components for the simplest ISP is only two, and the corresponding ISP is shown in Fig. 4.

In Fig. 4, it assumes that the image sensor produces digital R, G, and B data for each pixel. Unfortunately, no image sensors produce the R, G, and B data altogether at the same pixel position, unlike display devices where three or four color sub-pixels exist within a pixel. Foveon [5] invented the image sensor that samples R, G, and B data altogether at any pixel position. However, the practical sensor

Table 1 Timing reference code configuration

Data bit number	First word (FF)	Second word (00)	Third word (00)	Fourth word (XY)
7 (MSB)	1	0	0	1
6	1	0	0	F
5	1	0	0	V
4	1	0	0	H
3	1	0	0	P3
2	1	0	0	P2
1	1	0	0	P1
0	1	0	0	P0

Table 2 Protection bits in SAV and EAV

F	V	H	P3	P2	P1	P0
0	0	0	0	0	0	0
0	0	1	1	1	0	1
0	1	0	1	0	1	1
0	1	1	0	1	1	0
1	0	0	0	1	1	1
1	0	1	1	0	1	0
1	1	0	1	1	0	0
1	1	1	0	0	0	1

F = 0 during field 1; 1 during field 2
V = 0 elsewhere; 1 during field blanking
H = 0 in SAV; 1 in EAV
P0, P1, P2, P3: protection bits (see Table 2)

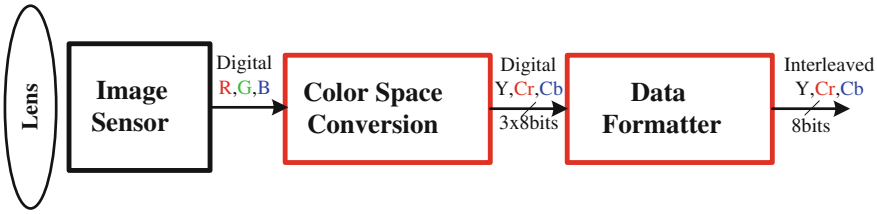


Fig. 4 Simplest ISP architecture

samples one of color components for a pixel as shown in Fig. 5 [6]. This differentiation between two light-related devices comes from the fact that the allowable pixel sizes they use are quite different.

Let's compare the size of two different optical devices shown in Fig. 6: a display panel and an image sensor supporting the same FHD (Full High Definition, 1920 × 1080) resolution, assuming the size of the display panel is 5 inches, and that of the image sensor is 1/3 inch. The display panel is easier to implement, compared to the image sensor because the effective pixel area of the display panel is $(3 \times 5)^2$ times as large as that of the image sensor. The pixel area of an image sensor needs to be as large as possible for improving SNR (Signal to Noise ratio). However, there is another constraint on the pixel size, which claims that an image sensor should be made as small as possible such that it can be packaged inside a compact smartphone of a small form factor. As the sensor shrinks, the SNR becomes lower. So there must be some compromise between the pixel size and the image sensor resolution. If sub-pixels constituting a pixel, e.g., *R*, *G*, *B* sub-pixels, are to be defined as in the display panel, the effective area of each sub-pixel will be much smaller. This is why sub-pixels of a pixel cannot be implemented on the same plane in the image sensor. Thus, the appropriate compromise between the high SNR and the small form factor is to adapt the spatial subsampling strategy such as Bayer array.

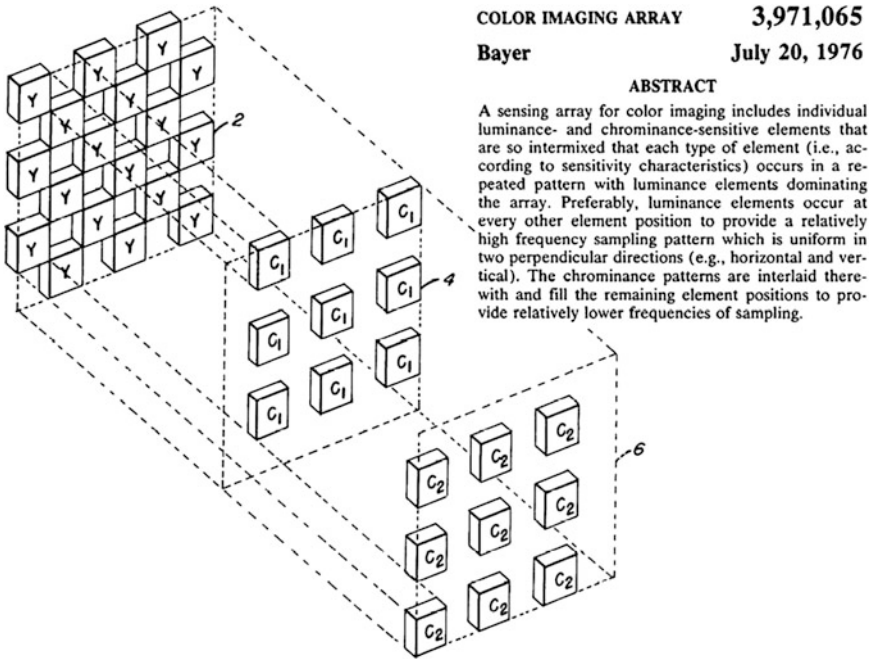


Fig. 5 Color imaging array by Bayer

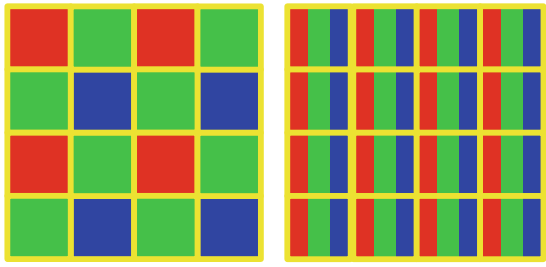


Fig. 6 Typical color filter pattern in image sensors (*left*) and display panels (*right*)

A sensor array made by spatial color subsampling is called color filter array (CFA) or Bayer array. According to the patent by Bayer, only the principle of color subsampling is provided, and which color is sampled is not explained. Thus, colors can be sampled in a variety of ways, and these combinations of sampling are also called Bayer pattern. Some typical examples of Bayer patterns are shown in Fig. 7.

Image sensors with Bayer pattern have high sensitivity with low implementation cost, but the process of restoring deficient color components is additionally required. This process is called interpolation or demosaicing. There are lots of ways [7] in demosaicing color filter arrays. One of the simplest is the 1-st order interpolation, i.e., bilinear interpolation. Among Bayer patterns mentioned in Fig. 7, the

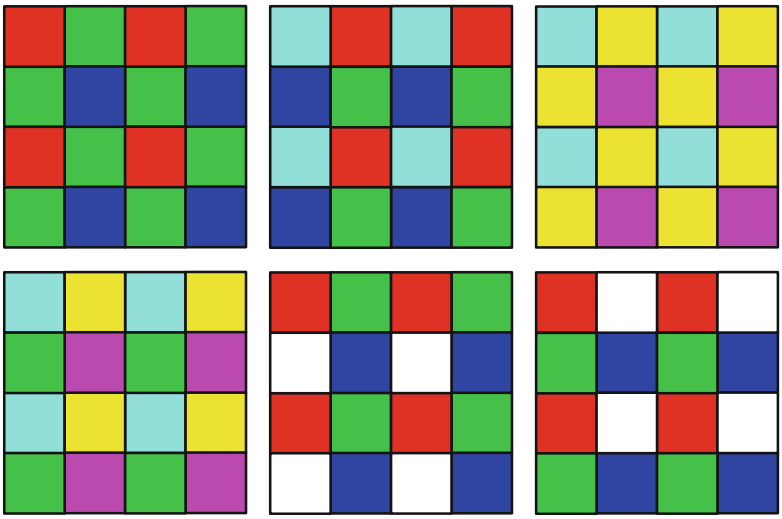


Fig. 7 Bayer patterns

RGB pattern is widely used since it allows better color reproduction. So, the subsidiaries of bilinear interpolation are to be described with this pattern. In Fig. 8, the shaded pixels represent practically sampled ones in the sensor array, while the other unshaded pixels represent those to be restored by bilinear interpolation in Eq. (3).

$$\begin{aligned} R11 &= R11 \\ R12 &= \frac{R11 + R13}{2} \\ R21 &= \frac{R11 + R31}{2} \\ R22 &= \frac{R11 + R13 + R31 + R33}{4} \end{aligned} \tag{3a}$$

$$\begin{aligned} G22 &= \frac{G12 + G21 + G23 + G32}{4} \\ G23 &= G23 \\ G32 &= G32 \\ G33 &= \frac{G23 + G32 + G34 + G43}{4} \end{aligned} \tag{3b}$$

$$
\begin{aligned}
B22 &= B22 \\
B23 &= \frac{B22 + B24}{2} \\
B32 &= \frac{B22 + B42}{2} \\
B33 &= \frac{B22 + B24 + B42 + B44}{4}
\end{aligned}
\tag{3c}
$$

Because bilinear interpolation averages two or four adjacent data of the same color attribute, the interpolated values may be what do not exist in the real scene. Since different interpolation equations are used for color components, the associated color built by combining them can show undesirable color where there are edges with high gradient. These unwanted color artifacts are called pseudo-color or color noise. Figure 9 shows artifacts produced by bilinear interpolation. The periodic noise pattern, which is called zipper noise (or maze noise), is shown with additive pseudo-color. The main role of color interpolation is to suppress such pseudo-color and zipper noise. The zipper noise can be reduced greatly by interpolating pixels along the distinct edges as shown in Fig. 9c.

Edge-directed interpolation is an adaptive approach, where the adjacent pixels around each pixel are analyzed to decide if there exists a horizontal or vertical edge. There are lots of ways to decide the direction of edges. In [19], the simplest form of edge direction detection is presented. Let $G22$ be interpolated using its neighboring

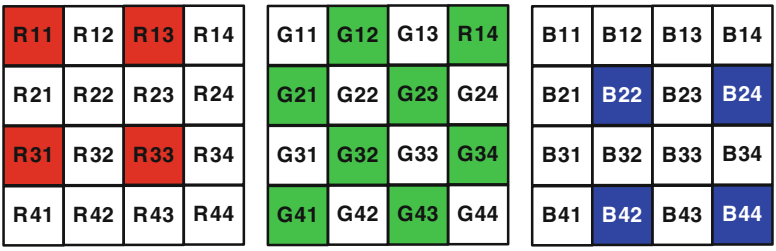


Fig. 8 Pixels to be interpolated by bilinear interpolation

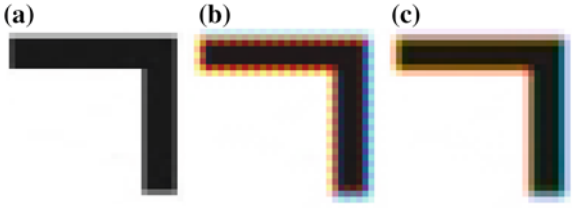


Fig. 9 Artifacts by color filter interpolation. **a** Original; **b** bilinear interpolation; **c** edge-directed interpolation [19]

G pixels in Fig. 8. The horizontal and vertical gradients are defined as $\Delta H = |G21 - G23|$ and $\Delta V = |G12 - G32|$ respectively. If $\Delta H > \Delta V$, the edge direction is vertical, then $G22 = (G12 + G32) \gg 1$. If $\Delta H < \Delta V$, the edge direction is horizontal, then $G22 = (G21 + G23) \gg 1$. Otherwise, $G22 = (G12 + G21 + G23 + G32) \gg 2$. In this way, the G image is interpolated first, and then the other color planes are acquired by utilizing the G image.

Impulsive noise is easy to remove by legacy noise reduction filters that utilize median filter. The basic assumption about noise is that noise is statistically independent and has very high-frequency components. Zipper noise looks like high-frequency noise, but it is difficult to remove by a legacy noise reduction filter because its frequency components are in the mid-to-high ranges. This is a contradiction to the basic assumption on noise. In Fig. 10 filtering results are presented by applying median filter and mean filter to remove zipper noise. The results show that the zipper noise is very difficult to remove by basic noise reduction tools. Thus, it is desirable to suppress the zipper noise in the interpolation stage instead of using noise reduction filter after color interpolation. Besides, an additional filter for removing pseudo-color is also necessary because it is hard to remove pseudo-color only with interpolation. Figure 9c also shows pseudo-colors along the edges after edge-directed interpolation

Figure 11 is the block diagram of an ISP evolved to compensate for artifacts raised by using a Bayer sensor. Anti-aliasing filter means a low-pass filter adapted before the color sampling to avoid aliasing. The ideal anti-aliasing filter must be an optical low-pass filter (OLPF) because the signals before the spatial subsampling are purely optical. However, OLPF cannot be considered here because it must be considered during the camera module design stage. Nevertheless, the first function of an ISP needs to be a noise reduction filter in the Bayer domain. The purpose of placing the noise filter here is not to prevent aliasing, but to prevent noise propagation through color interpolation. Anyway, this function is often called anti-aliasing noise filter for convenience. By adopting a cost-effective Bayer sensor, thus, an ISP should add following functions as shown in Fig. 11. Among them, the CFA interpolation is mandatory and all noise filters are optional.

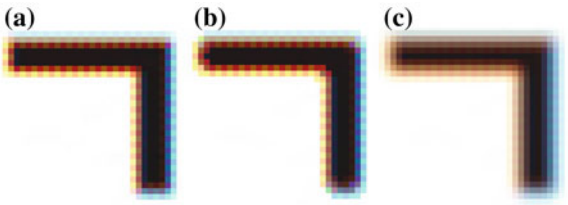


Fig. 10 Applying conventional low-pass filter to reduce zipper noise. **a** zipper noise; **b** median filtering; **c** mean filter

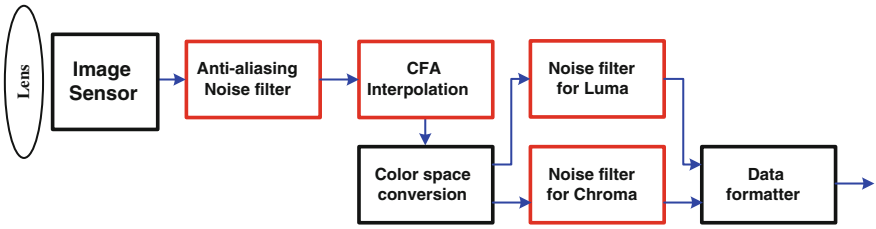


Fig. 11 ISP architecture to recover artifacts from a Bayer image sensor

2.1 Anti-aliasing Noise Filter

Noise needs to be removed in the Bayer domain. Salt-and-pepper noise produced during the manufacturing of image sensor has to be removed before color interpolation. Otherwise the noise will be expanded through color interpolation kernel.

2.2 Color Filter Array Interpolation

This is the process to restore the original color components from the sampled ones. It results in zipper noise and pseudo-color. The zipper noise can be suppressed considering edge direction during color interpolation process.

2.3 Noise Filter for Luma

In an anti-aliasing noise filter, it is not possible to exploit correlation with the adjacent pixels because they are of different color attributes. After interpolation, it is easier to remove Gaussian noise by considering correlation with adjacent data. This noise filter is a legacy noise filter [8] that has been developed for a long time.

2.4 Noise Filter for Chrominance: C_B *and* C_R

This is a filter for removing pseudo-color caused by subsampling and interpolation process. Because human eyes are very sensitive to rapid color changes, it is necessary to build a natural image by suppressing excessive color changes.

3 ISP Architecture for Color Reproduction

The process for restoring 'natural' color is necessary because the response of silicon to light is quite different from that of human eyes. Color is the response of light receptors of the human eyes to light spectrum. In the retina where there are rods and cones, rods sense brightness and cones sense chromaticity, respectively. Rods are extremely sensitive to light, and can be triggered by as few as six photons [9]. At very low light conditions, visual experience is solely decided by rods. Cones require significantly brighter light than rods. There are three different types of cones, distinguished by their response pattern with different wavelengths of light. Colors can be defined and quantified by the degree with which these cells are stimulated.

A color space is a 3-dimensional representation system into which a perceived color is translated [10]. The whole colors are represented by three-dimensional coordinates in a color space. There are many color spaces such as CIERGB, CIEXYZ, CIELAB, CIELUV, and so on. An RGB color space [11] is any additive color space based on the RGB color model. The most popular RGB color space is sRGB [12], which is used in consumer electronics including digital cameras, video cameras, televisions, projectors, and computer monitors. The RGB color space is defined in Rec. ITU-R. BT. 709.

A particular RGB color space is defined by the three chromaticities of the red, green, and blue additive primaries, and can produce any chromaticity inside the triangle whose vertices are defined by those primary colors. The primary colors are specified with reference to their corresponding chromaticity coordinates (x, y) in the CIE 1931 color space [13]. To completely specify an RGB color space, a white point and a gamma correction curve need to be additionally defined. In Table 3, the three primary colors and white points for popular RGB color spaces are summarized [11].

It should be noted that a gamma correction curve is mandatorily included for specifying a color space. Our "nonlinear" eyes do not perceive light like "linear" image sensors. They are more sensitive to changes in dark tones, and less in bright

Table 3 RGB color space parameters

Color space	Gamut	White point	Primaries					
			Red		Green		Blue	
			x	y	x	y	x	y
sRGB, HDTV	CRT	D65	0.64	0.33	0.30	0.60	0.15	0.06
PAL/SECAM	CRT	D65	0.64	0.33	0.29	0.60	0.15	0.06
NTSC(1987)	CRT	D65	0.63	0.34	0.31	0.595	0.155	0.07
UHDTV	Wide	D65	0.708	0.292	0.170	0.797	0.131	0.046
CIE(1931) RGB	Wide	E	0.7347	0.2653	0.2738	0.7174	0.1666	0.0089

tones, compared to silicon image sensor. It is because human eyes have evolved to enable our vision system to operate over a wide range of luminance. Gamma correction or gamma encoding is the name of a nonlinear operation used to code and decode luminance or tri-stimulus values in image sensors or display systems. Gamma correction is defined by the following power-law expression:

$$V_o = V_i^{\gamma} \tag{4}$$

The input and output values are nonnegative real numbers and are typically in the range of [0,1]. A gamma value which is smaller than 1 (i.e. $\gamma < 1$) is called an encoding gamma, and is used to compress the dynamic range of input values. The nonlinear characteristics of the human eyes to the brightness change can be observed from the exemplary patches in Fig. 12. Figure 12 shows what happens after quantizing continuous tones in an explicitly linear way or a perceptually linear way. Figure 12b shows quantizing into 32 levels by uniform quantization step and Fig. 12c by nonlinear quantization step based on a gamma curve. The quantization step size in Fig. 12c is numerically nonlinear but is perceived linear to the human eyes, while the quantization step size in Fig. 12b is arithmetically linear but looks nonlinear. Thus the true linear response of an image sensor should be perceived linear for human eyes. The nonlinear tone mapping process for human eyes is called gamma correction, which is included in defining a color space. The definition of gamma curve in the sRGB color space is like Eq. (5) below.

$$V_o = \begin{cases} 1.099 V_i^{0.45} - 0.099, & 0.018 \le V_i \le 1 \\ 4.500 V_i, & 0 \le V_i < 0.018 \end{cases} \tag{5}$$

To support a particular RGB color space, both tone mapping and color mapping are required. The former is gamma correction and the latter is color correction. Gamma correction is nonlinear but color correction is linear, which are generally implemented by a 3 × 3 matrix multiplication. In color correction, the result of color mapping should not be affected by the brightness level of the captured scene. To

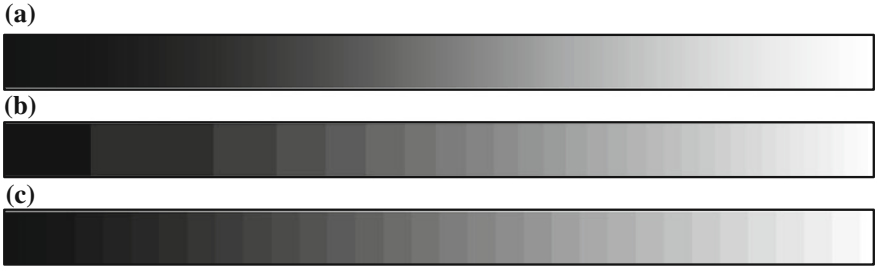


Fig. 12 Linear and nonlinear quantization of continuous tones. **a** Continuous tones from 0 to 1023; **b** linearly quantized tones into 32-levels; **c** nonlinearly quantized tones into 32-levels according to a gamma curve

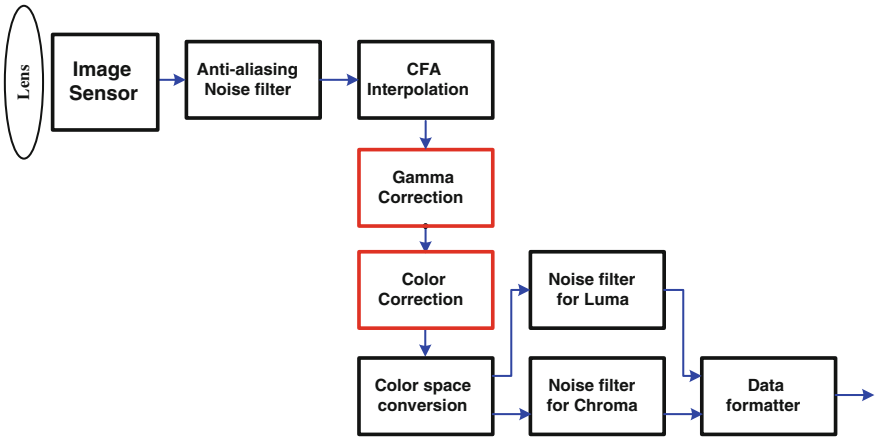


Fig. 13 ISP architecture to support a particular RGB color space

maintain consistency of color correction regardless of brightness, the color correction process needs to be linear. An ISP pipeline containing these two functions to support a particular color space is depicted in Fig. 13.

There is no restriction as to where stage gamma correction is placed. Gamma correction can be located before color interpolation or after color correction. It is also possible to place it even after color space conversion. The purpose of doing gamma correction is a nonlinear tone mapping. As long as this purpose is achieved efficiently, the place of performing gamma correction in the ISP pipeline is not so important. In reality, many ISP implementers do not use gamma correction in the same way. If efficient hardware implementation is pursued, implementing gamma correction in the Bayer domain or in the Y-C_B-C_R domain may be more efficient than in the RGB domain. Figure 14 shows two modified ISP chains with different location for gamma correction.

A white point is also included in the definition of an RGB color space. It is used to standardize the light spectrum and is abbreviated as D65 or E as tabulated in Table 3. The spectrum of a standard illuminant can be converted into tri-stimulus values by integrating it over all wavelength spectrums. The set of resultant three tri-stimulus coordinates of an illuminant is called a white point. CIE Standard Illuminant D65 [14] is a commonly used standard illuminant defined by the CIE. It describes standard illumination conditions at open-air in different parts of the world. D65 is intended to represent average daylight, and has a corresponding color temperature of approximately 6500 K. The power spectrum of illuminant D65 is shown in Fig. 15. CIE standard illuminant D65 should be used in all colorimetric calculations requiring representative daylight, unless there are specific reasons for using different illuminant. Illuminant E [15] is an equal-radiator; it has a constant distribution inside the visible spectrum. That is, it is a theoretical illuminant that gives equal weight to all wavelengths.

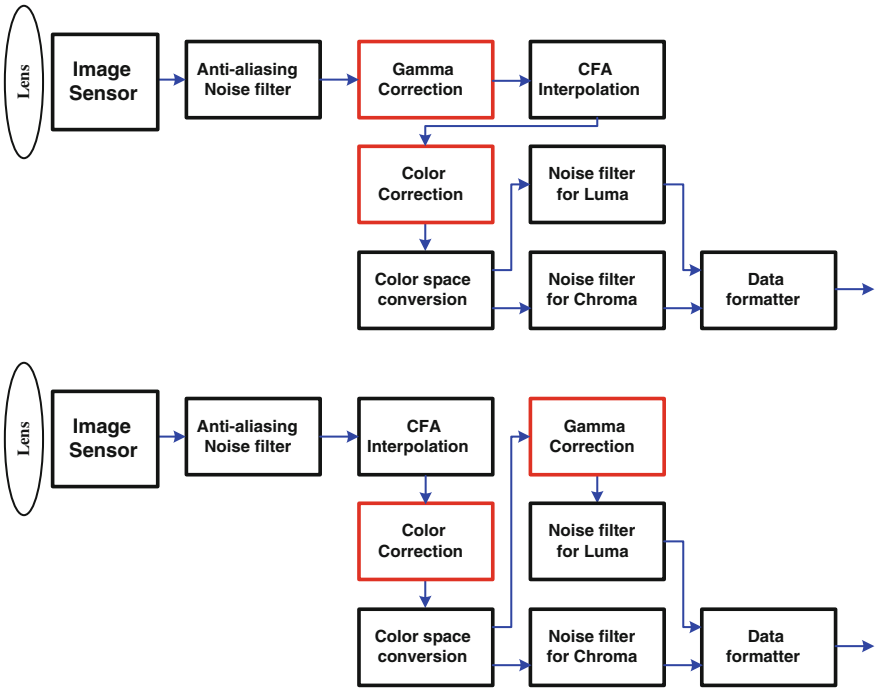


Fig. 14 ISP architecture variants for gamma correction. **a** ISP architecture variant with gamma correction at the Bayer domain; **b** ISP architecture variant with gamma correction at the Y-C_B-C_R domain

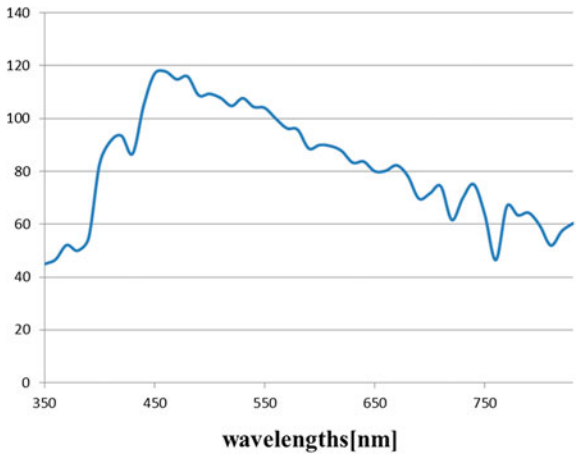


Fig. 15 Spectral power distribution of Illuminant D65

Color correction is performed with reference to the illumination spectrum of D65. If the illumination spectrum is different from D65, the chromaticity of the same object in the same scene will be perceived different in colors. Thus the chromaticity for the current light spectrum has to be corrected to be perceived similar to that of D65, since the chromaticity under D65 is most natural to average people. The attribute of light source is numerically characterized by the color temperature. Thus, the current color temperature should be changed to match D65. This process is called AWB (Auto-White Balance) and it is to compensate for the color distortion caused by the light spectrum different from D65. The key technology of AWB is to measure the color temperature of the current light source. To do this, achromatic-colored regions in the scene are used to estimate the color temperature because the color there reflects the color temperature of the light source. Gray or white regions are typical achromatic-colored regions. Achromatic color region is where the ratios between R, G, and B components are identical. Thus, the AWB process is modeled by Eq. (6), where average values of R, G, and B components in achromatic-colored region are denoted as $\bar{R}, \bar{G}, \bar{B}$ respectively. It is desirable for AWB to be performed before color correction.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} \bar{G}/\bar{R} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \bar{G}/\bar{B} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6}$$

However, locating achromatic-colored region is almost impossible in practice; even human eyes sometimes cannot identify it from the natural scene. However, it is possible to measure the color variations of achromatic-colored region when the color temperature of the ambient light is changed. Such color variation is confined in a small area of the color gamut. Such area can be identified experimentally in the C_B-C_R plane by plotting the C_B-C_R components of achromatic-colored regions for all allowable color temperatures, as shown in Fig. 16. Then we can find connected regions where at least one achromatic-colored pixel exists. Among them, we can assume that there exists one region which reflects the current ambient color temperature. In this way, we can estimate the ambient color temperature. There are many heuristic ways to decide which area gives a good estimate for the ambient

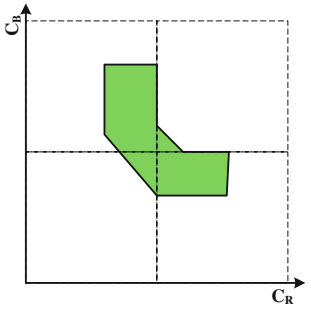


Fig. 16 Determination of the chrominance variation of achromatic-colored regions in the C_B-C_R plane

color temperature. Besides, instead of using C_B/C_R components, other terms can be used such as *G*-*R* and *G*-*B*, *G*/*B* and *G*/*R*, and so on.

Chromaticity is an objective specification of a color regardless of its brightness, and is further represented by hue and saturation. The white point is a neutral reference, which is characterized by chromaticity. All other chromaticities are defined with respect to the white point using polar coordinates (an angle and the distance from the origin). Hue is "the degree to which a stimulus can be described as similar to or different from stimuli that are described as red, green, and blue" [16].

HSL (Hue-Saturation-Lightness) and HSV (Hue-Saturation-Value) are the two most common cylindrical-coordinate representations of points in an RGB color space [17]. Figure 17 shows the hue from 0° to 360°. In case of emphasizing or deemphasizing particular color, we first find the hue corresponding to that color and then emphasize or deemphasize all *R*, *G*, and *B* values that have the same hue. In case of adjusting color of the whole image consistently, it is done by rotating the hue of each RGB data around the white point as much as needed.

Calculating the hue from *R*, *G*, and *B* data requires very complicated operation. So the *Y*-C_R-C_B color space can be regarded as a cost-effective substitute for hue. Mapping all colors in the C_R-C_B plane is shown in Fig. 18. Hue control in the C_R-C_B plane can be performed by Eq. (7a). The constant 128 indicates that the values of *Y*-C_R-C_B are in 8-bit, which is replaced by 512 when using 10-bit data. Saturation control is the same as amplifying the C_B and C_R components according to Eq. (7b). It is performed by Eq. (7c) when both hue and saturation controls are conducted simultaneously.

$$\begin{bmatrix} C'_B - 128 \\ C'_R - 128 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} C_B - 128 \\ C_R - 128 \end{bmatrix} \tag{7a}$$

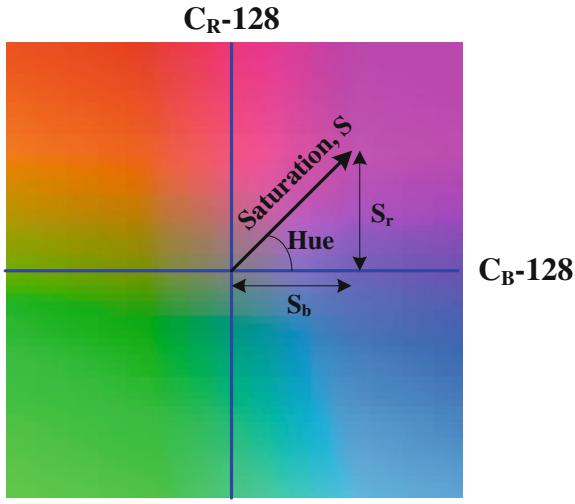
$$\begin{bmatrix} C'_B - 128 \\ C'_R - 128 \end{bmatrix} = \begin{bmatrix} S_b & 0 \\ 0 & S_r \end{bmatrix} \begin{bmatrix} C_B - 128 \\ C_R - 128 \end{bmatrix} \tag{7b}$$

$$\begin{aligned} \begin{bmatrix} C'_B - 128 \\ C'_R - 128 \end{bmatrix} &= \begin{bmatrix} S_b & 0 \\ 0 & S_r \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} C_B - 128 \\ C_R - 128 \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} S_b & 0 \\ 0 & S_r \end{bmatrix} \begin{bmatrix} C_B - 128 \\ C_R - 128 \end{bmatrix} \\ &= \begin{bmatrix} S_b\cos\theta & -S_b\sin\theta \\ S_r\sin\theta & S_r\cos\theta \end{bmatrix} \begin{bmatrix} C_B - 128 \\ C_R - 128 \end{bmatrix} \end{aligned} \tag{7c}$$

Various functions are included in an ISP for reproducing correct colors that human eyes perceive, and are performed in different color domains. AWB is

Fig. 17 Hue in the HSB/HSL encodings of RGB

Fig. 18 Color distribution in the C_R-C_B plane



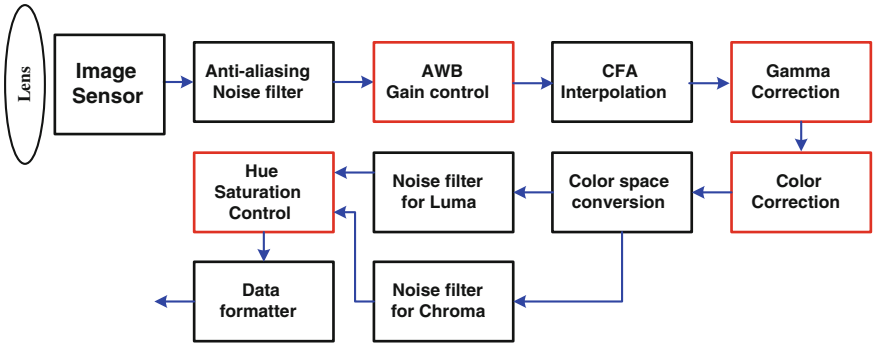


Fig. 19 ISP architecture for color reproduction

performed usually in the Bayer domain, both gamma correction and color correction are done in the RGB domain, and the hue/saturation control is conducted in the Y-C_R-C_B domain (See Fig. 19).

4 ISP Architecture with Pre-/Post-processing

Some additional pre-/post-processing functions are added to the baseline ISP pipeline described above. The purpose of pre-processing is to compensate for the sensor or camera distortions such that robust images can be acquired through a legacy ISP pipeline. The role of post-processing is to give a better visual quality from the standpoint of human visual system.

An image sensor has permanent bright or dark pixels due to physical defects. They are called dead or defective pixels, and the function to remove them is called DPC (Deal Pixel Concealment). Dead pixels are far brighter and much darker than their neighbors and generate salt-and-pepper noise. They can be easily removed by using a median filter, which always leads to a blurred image. A special noise filter has been developed to effectively remove the salt-and-pepper noise. This filter detects the dead pixels in real time and corrects them by replacing them with neighbor pixel data. Another method is to correct predefined dead pixels whose locations are searched and stored in the memory in advance. This method is free from the risk of blurring by a legacy noise filter because it conceals only predetermined defect pixels. The more coordinates of defective pixels are stored, the more high-cost memory is consumed. So, an appropriate memory storage should be determined in terms of cost and performance.

Sensor response does not have perfect linearity. Each pixel of an image sensor is a capacitive photodiode, and the charge in each pixel is discharged according to the incident photons. The discharged charge is sampled in voltage, and is regarded as a pixel value. Naturally it is not possible to detect no-light condition because the photodiode is always discharged by the reverse bias current, even in no-light condition. To detect sensor response corresponding to no-light, any image sensor has the dedicated sensor region called optical black area. The optical black area has the same structure as that of normal pixels, but it is made intentionally not to be exposed to light by covering photo-diodes with metal. Thus, it is possible to estimate the sensor response at no-light condition. Because there are R, G, and B pixels in the optical black area, it is possible to have sensor responses to '0' in no-light condition if their averaged values in the optical black area are subtracted from the sensor output appropriately. The function to do this is called BLC (Black Level Compensation) and is implemented by using Eq. (8), where OB_R, OB_G, and OB_B are the average values of red, green, and blue pixels in the optical black area, respectively. BLC should be the function to operate at the earliest stage in an ISP pipeline because only this function can make the sensor response linear.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} R \\ G \\ B \end{bmatrix} - \begin{bmatrix} OB_R \\ OB_G \\ OB_B \end{bmatrix} \tag{8}$$

The magnitude and brightness of each pixel will have linearity after BLC. However, the linear slope of each pixel is not constant but varies randomly

Fig. 20 Flat-field image without lens-shading correction

according to its spatial position. The image is brightest in the center of optical axis and becomes monotonically darker as one goes to the edge of the field-of-view. The shading might be caused by nonuniform illumination or nonuniform camera sensitivity. In general this shading effect is mainly due to a lens system, and is called lens-shading distortion. Figure 20 shows a lens-shading image which is an originally flat-field image having a constant value all over the plane.

LSC (Lens-Shading Correction) is the process to compensate for the disparity of linear gain of each pixel due to lens shading, such that all pixels can have the same light-to-voltage gain regardless of their locations in the sensor array. The simplest and robust solution for LSC is to compensate for shading by the correction gain, which was estimated for each pixel in advance and then stored in the memory. This method is called FFC (Flat Field Compensation) [20]. FFC consists of two numbers for each pixel, the pixel's gain and its dark current. The corrected image $C(x, y)$ at the pixel location (x, y) is obtained by Eq. (9).

$$C(x,y) = \frac{R(x,y) - D(x,y)}{F(x,y) - D(x,y)} \cdot m \tag{9}$$

where, $D(x, y)$ is a dark frame, $R(x, y)$ is a raw image, $F(x, y)$ is a flat-field image, and m is the average value of $F(x, y)$–$D(x, y)$. The dark frame and the flat field are captured experimentally by taking the flat-field scenes in a very dark lighting condition and in a marginally unsaturated lighting condition. FFC is not appropriate in a baseline ISP because it requires sufficiently large memory to store the entire image. Instead, an appropriate mathematic model for the LSC gain map is used.

Noise reduction is performed after consistent linearity is obtained for the whole pixels. Noise sources in an image are various. The need for noise reduction is increasing as the resolution of image sensor is increased with the pixel size being drastically reduced. Noise reduction is considered as a key component to determine the performance of camera systems, and consumes the most computational power of legacy ISPs.

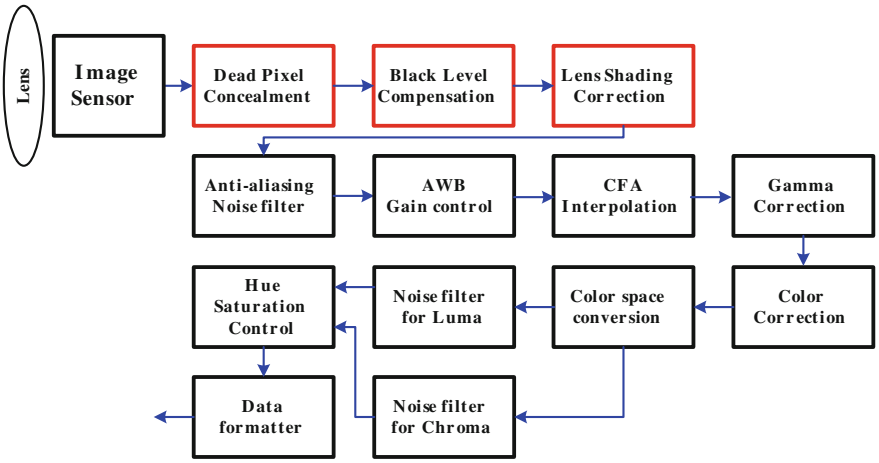


Fig. 21 ISP architecture for handling sensor derating factors

The enumerated methods so far are the functions to let the sensor response to be linear and to compensate for derating factors of an image sensor or a camera module. These are not mandatory functions in a baseline ISP, but need to be considered whether to implement or not, since they try to compensate for the imperfection of a camera system. Figure 21 shows an ISP pipeline with such compensation functions. LSC should be located after BLC, but there is no strict restriction on DPC location if and only if the DPC is located before color interpolation in a baseline ISP chain. Thus it is often desirable to embed DPC function in anti-aliasing noise filter.

Let's examine typical methods to enhance subjective visual quality. Mach bands [18] are an optical illusion, which can be seen in an image patch where there are two wide bands, one light and the other dark, separated by a narrow strip with a light-to-dark gradient. Human eyes perceive two narrow bands of different brightness at either side of the gradient that are not present in the original image (See Fig. 22).

Fig. 22 Mach bands as optical illusion

Edge enhancement is a digital processing technique to improve the sharpness of an image by intentionally emphasizing Mach band effect. The creation of bright and dark highlights on either side of any line makes the line look contrasted from a distance. It only increases the perceptual sharpness. Some artifacts are raised by edge enhancement. The enhancement is not completely reversible, and some detail in the image can be lost as a result of enhancement. Repeated sharpening operations on the resulting image compound the loss of detail, and lead to artifacts known as ringing. Most sharpening filters are based on the first and the second-order derivatives. Among them, Laplacian filter has been the most popular tool. Equation (10) describes one of the Laplacian filters for the pixel value $I(x, y)$, where x and y are horizontal and vertical coordinates in an image.

$$\begin{aligned} L(x,y) &= \nabla^2(x,y) = \frac{\partial^2 I(x,y)}{\partial x^2} + \frac{\partial^2 I(x,y)}{\partial y^2} \\ &= I(x-1,y) + I(x+1,y) + I(x,y-1) + I(x,y+1) - 4I(x,y) \end{aligned} \tag{10}$$

Contrast is the difference in color and light that makes an object distinguishable from others and the background. The human visual system is more sensitive to contrast than absolute luminance. The contrast-controlled value is acquired by Eq. (11), where K_c, K_r and K_b are contrast control gain, reference luminance, brightness control offset, respectively. The contrast gain K_c is a fractional number ranging from 0 to 1. The reference luminance K_r is defined as 2^{B-1} if B-bit codes are used for luminance representation. The brightness offset K_b is used to increase or decrease the average brightness level.

$$Y' = K_c(Y - K_r) + K_b + K_r \tag{11}$$

Figure 23 shows the proposed baseline ISP pipeline considering all related standards. The proposed ISP chain will be the minimum configuration for designing a baseline ISP.

5 Further Works on ISP

The ISP itself is a pipelined chain of functional units, whose inputs are fed from the previous unit and the processed outputs are transferred to the next unit. In each functional unit, every pixel of an image is processed sequentially. When a pixel is processed, only its adjacent pixels are utilized and a small window is defined around the pixel such that some lines of the incoming image have to be stored in the memory. For defining an $N \times N$ window, at least $N - 1$ lines are to be saved into the memory. In other words, it is often said that $N - 1$ line memories are required. In this way, an ISP only utilizes the spatially localized information. One of the hot functions in a legacy ISP is mainly focused on the true color reproduction. As the

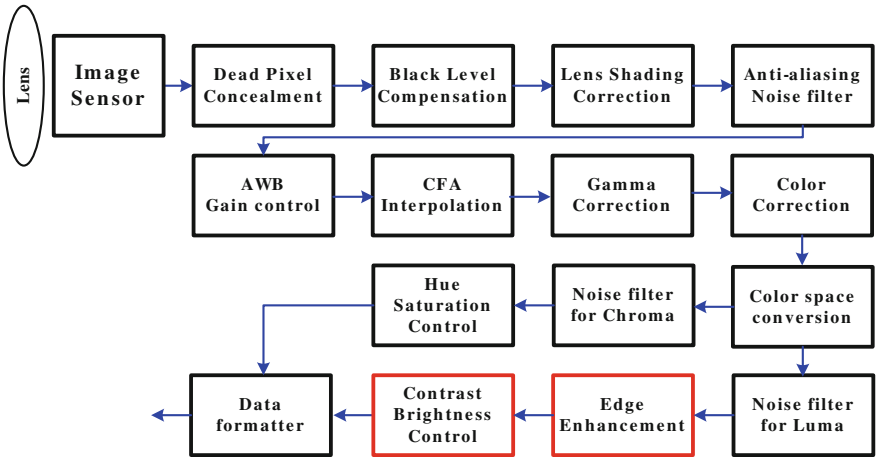


Fig. 23 Proposed baseline ISP pipeline

ambient color temperature changes, the color-related functions begin to degrade the subjective color quality. Realizing the robust color quality over the ambient color temperature is becoming a critical requirement for high-quality ISP implementation. It is because the drastic change of chrominance is annoying to human eyes while the drastic change of luminance is perceived as natural and often can be ignored without annoying our eyes.

When global information is necessary, an entire image has to be saved in the memory. In this case, the amount of memory requirement is so huge such that the frame memory is realized by using an external SDRAM. When the frame memory is available, a more sophisticated function can be conducted in software by using a powerful CPU and/or a GPGPU (General Purpose Graphic Processing Unit). Nowadays, many computer vision applications have been implemented in an intelligent camera. In a legacy ISP, however, those functions requiring the frame memory are not considered since they cannot be embedded inside an image sensor. Thus, they are not regarded as further works for ISP. They are highly related to the intelligent camera. Recently many researches have been done to expand the dynamic range of the image sensor. WDR (Wide Dynamic Range) or HDR (High Dynamic Range) implies such a technique to expand the dynamic range by utilizing two or more frames respectively and is not considered in a legacy ISP pipeline since it requires the frame memory.

There remain a few functions to be implemented in an ISP. Nevertheless, color interpolation and noise reduction are always key functions that need more improvement. Besides, the false color suppression or pseudo-color removal is also becoming a major function since the false color critically distorts the human eyes.

Acknowledgments This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Recommendation ITU-R BT.601-7, Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios, ITU, March 2011
2. Recommendation ITU-R BT.656-4, Interface for digital component video signals in 525-line and 625-line television systems operating at the 4:2:2 level of Recommendation ITU-R BT.601, ITU, Dec 2007
3. Recommendation ITU-R BT.709-5, Parameter values for the HDTV standards for production and international programme exchange, ITU, Feb 2004
4. Recommendation ITU-R BT.2020-1, Parameter values for ultra-high definition television systems for production and international programme exchange, ITU, July 2014
5. http://en.wikipedia.org/wiki/Foveon_X3_sensor
6. Bayer BE (1976) US Patent 3971065. Color imaging array. Accessed 20 July 1976
7. Gunturk Bahadir K, Glotzbach John, Altunbasak Yucel, Schafer Ronald W, Mersereau Russel M (2005) Demosaicking: color filter array interpolation. IEEE Signal Process Mag 22 (1):44–54
8. Buades A, Coll B, Morel JM (2005) A review of image denoising algorithms, with a new one. Multisc Model Simul 4(2):490–530
9. http://en.wikipedia.org/wiki/Retina
10. http://en.wikipedia.org/wiki/Color_space
11. http://en.wikipedia.org/wiki/RGB_color_space
12. http://en.wikipedia.org/wiki/SRGB
13. http://en.wikipedia.org/wiki/CIE_1931_color_space
14. http://en.wikipedia.org/wiki/Illuminant_D65
15. http://en.wikipedia.org/wiki/Standard_illuminant#Illuminant_E
16. http://en.wikipedia.org/wiki/Hue
17. http://en.wikipedia.org/wiki/HSL_and_HSV
18. http://en.wikipedia.org/wiki/Mach_bands
19. Laroche CA, Prescott MA (1994) Apparatus and method for adaptively interpolating a full color image utilizing chrominance gradients. US Patent 5,373,322
20. http://en.wikipedia.org/wiki/Flat-field_correction

Springer

http://www.springer.com/978-94-017-9986-7

Theory and Applications of Smart Cameras
Kyung, C.-M. (Ed.)
2016, VI, 366 p. 202 illus., 135 illus. in color., Hardcover
ISBN: 978-94-017-9986-7