IEEE Open Journal of
Circuits and Systems

Received 23 June 2022; revised 15 September 2022; accepted 26 September 2022. Date of current version 24 November 2022.

Digital Object Identifier 10.1109/OJCAS.2022.3213062

A Vision System With 1-inch 17-Mpixel 1000-fps Block-Controlled Coded-Exposure Stacked-CMOS Image Sensor for Computational Imaging and Adaptive Dynamic Range Control

TOMOKI HIRATA, HIRONOBU MURATA, TAKU ARII, HIDEAKI MATSUDA, HAJIME YONEMOCHI, YOJIRO TEZUKA, AND SHIRO TSUNAI

Opto Device Development Center, Advanced Technology Research and Development Division, Nikon Corporation, Tokyo 108-6290, Japan

This article was recommended by Associate Editor P.-H. Hsieh.

CORRESPONDING AUTHOR: T. HIRATA (e-mail: tomoki.hirata@nikon.com)

ABSTRACT This study introduces a vision system that can acquire images at high speeds and high resolutions. Image sensors are used not only in digital still cameras but also in various applications that require capturing wide luminance differences beyond human perception. For example, fast, high-resolution object recognition, and motion tracking in automatic driving systems are essential, particularly in dark tunnels or the mid-summer sunshine. However, the resolution, frame rate, pixel size, and dynamic range should be traded off to achieve a high performance in capturing moving objects with a high contrast. We have developed a high-speed vision system with a readout operation of 1000 fps, resolution of 4K $\times$ 4K, dynamic range of 110 dB, and fine pixels of 2.7 μm. These characteristics were achieved using several technologies such as 1) coded exposure (CE), which divides the image plane into smaller blocks and controls the exposure time of each block individually, 2) arrangement of analog-to-digital converters in parallel for each block, and 3) three-dimensional wafer stacking, which enables high-density integration of circuits and pixels. The proposed system can be applied in CE-based computational imaging in addition to high-dynamic-range applications for handling both the dark and bright areas in a scene.

INDEX TERMS Block parallel, coded exposure (CE), high dynamic range (HDR), high-speed imaging, stacked CMOS image sensor (CIS), vision system.

I. INTRODUCTION

IMAGE sensors are not only used for taking photos but are also increasingly expected to serve as intelligent systems with surrounding configurations. Coded exposure (CE) [1], [2] is a method applied in intelligent system approaches; thus, various functions can be realized by selecting an integration variable in the plenoptic function [3]. A high dynamic range (HDR) can be realized when the integration variable is time. Various methods have been proposed for achieving HDR. The lateral overflow integration capacitor (LOFIC) method was introduced that could provide a plurality of detection capacitors [4]. Another method was presented to prevent photo diode (PD) saturation by adding low-sensitivity pixels [5]. However, the proposed methods required an enlarged pixel size. Alternatively, a high-speed readout, such as an array-parallel analog-to-digital converter (ADC) structure [6], is useful for integrating multiple frames [7] to realize HDR. However, this increases the noise level and requires a faster readout to reduce motion artifacts. To mitigate these adverse effects, a method was proposed in which a pixel array was divided into multiple blocks, and the signal integration time of each block was individually controlled [8]. In another method, CE was demonstrated using the pixel-level control of exposure time [9], [10], [11], [12], [13], [14], [15], [16] in addition to the HDR application. In these methods, the readout path and control circuitry should be arranged within the same plane because of unstacked sensors used; thus, the pixel

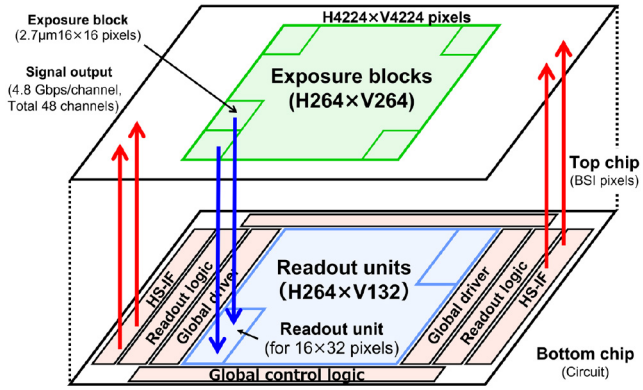


FIGURE 1. Device structure.

size is relatively large, and a high resolution is difficult to realize. Furthermore, a digital pixel sensor that used the stacked technology to expand DR was proposed [17]. Indeed, it measured the time until PD saturated for each pixel.

We have reported on a sensor architecture and system configuration that can achieve 4K × 4K resolution and 1000fps high-speed readout at the same time [18], [19]. In this paper, we describe the details of the vision system: key technologies for 1000 fps operation, signal processing and control algorithms, and experimental result of CE applications. We demonstrated the ability of CE by individually controlling the exposure time for each block of pixels using a stacked structure.

The remainder of this paper is organized as follows. In Section II, the architecture of the proposed image sensor is described. Section III presents the configuration of the experimental system using the image sensor and field-programmable gate arrays (FPGA). The implementation and performance results of the prototype image sensor are presented in Section IV. Section V discusses the experimental results for HDR imaging and other applications. Finally, Section VI draws the conclusions.

II. SENSOR ARCHITECTURE

A. BLOCK DIAGRAM

Fig. 1 shows a conceptual diagram of the proposed image sensor. This 1-inch image sensor has two layers: the top chip comprises back side illumination (BSI) pixels, and the bottom chip is used for signal processing; the layers are bonded to each other. The area with the highest density consists of two contacts per pixel. The top chip has a pixel array, which is divided into H264 × V264 exposure control blocks with a basic unit of H16 × V16 pixels. Smaller block size is better, but area is limited. The optimal solution for pixel size, circuit area, process, connection electrode density between 2 layers, and ease of signal processing is 16x16. The bottom chip has ADC circuits, logic circuits, high-speed interfaces (HS-IF), and H264 × V132 readout circuits (readout units), which are arranged directly below the pixels. Each readout unit corresponds to H16 × V32 pixels as the basic unit.

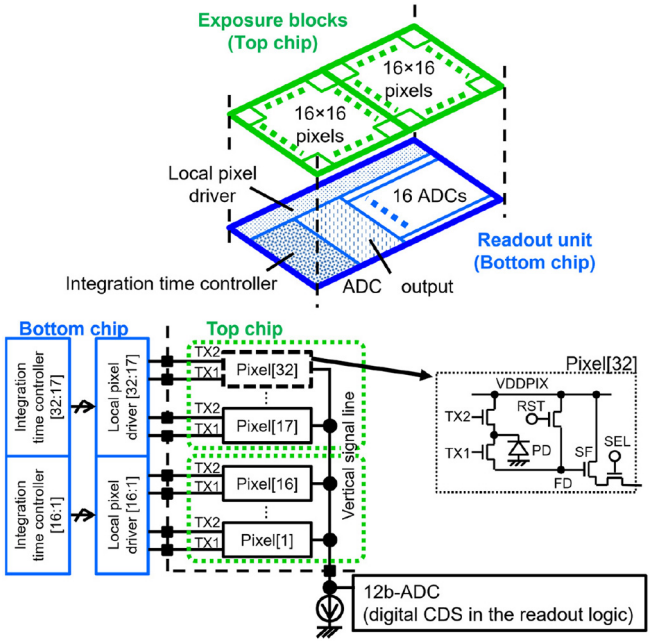


FIGURE 2. Exposure block and readout unit.

B. BLOCK ARRAY STRUCTURE

Fig. 2 shows a schematic of the readout unit and two exposure blocks. Each pixel comprised a PD and five transistors. The reset transistor (RST) and select transistor (SEL) were controlled by a global pixel driver placed in the periphery of the bottom chip. Both the photoelectron transfer transistor (TX1) and PD reset transistor (TX2) were controlled for each block by a local pixel driver in the readout unit. The readout unit is composed of 16 column ADCs, a data transfer circuit, an integration time controller, and a local pixel driver. Moreover, ADC is a single-slope type, and the ramp signal and counter are supplied by the peripheral circuitry of the bottom chip. Each ADC is connected to its corresponding pixels (1-32) via a vertical signal line and source followers (SFs).

C. KEY TECHNOLOGIES FOR 1000 FPS OPERATION

Fig. 3 shows a signal readout block diagram from the pixel to the HS-IF and a timing chart for one row period. Pixel signals were converted into digital data using ADC. Binary conversion and correlated double-sampling (CDS) operations were performed using the readout logic. Subsequently, signals were output through the HS-IF. Frame rate of this block parallel architecture can be defined as follows:

$$fps = \frac{1}{T_{row}} \times \frac{N_{v_{unit}}}{N_{v_{pix}}}$$
$$T_{row} = \left(T_{pix} + T_{vline} + T_{adc}\right) + T_{transfer} + T_{IF} \quad (1)$$

where T_{row}, $N_{v_{pix}}$, $N_{v_{unit}}$, T_{pix}, T_{vline}, T_{adc}, $T_{transfer}$ and T_{IF} are readout time per row, total vertical pixels, vertical number of readout unit, pixel reset and transfer time, settling time of vertical signal line (VLN), conversion time of ADC,

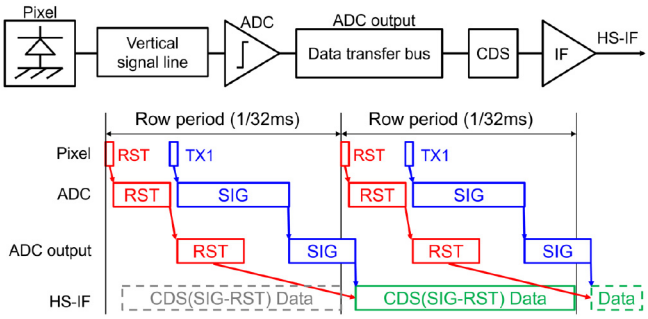


FIGURE 3. Signal readout flow.

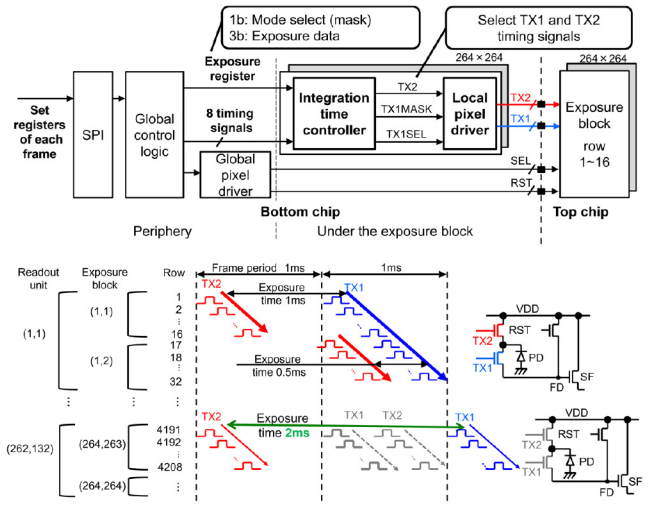


FIGURE 4. Block exposure control diagram.

data transfer time from ADC to peripheral readout logic and the time of stream out. This image sensor has three key technologies and methods for high-speed operation. The first is to reduce the settling time by dividing the VLN for each readout unit. The second is arranging ADC for each block to improve the parallel processing. The third is the pipeline operation of ADC, data transfer and HS-IF. In the current configuration, frame rate is limited by HS-IF to 1000fps.

D. BLOCK EXPOSURE CONTROL STRUCTURE

Operational modes of the exposure control include the "No Skip Mode" for short exposure times less than one frame, and "Skip Mode" for long exposure times more than one frame. Fig. 4 shows a simplified block diagram and timing chart of the exposure control. The integration time controller comprises a row counter, an address decoder, and four flip-flops that hold an exposure time register setting. The register has four bits per block, of which three are used to encode eight exposure times within one frame. The remaining bit is a mode-select signal, which is used as a mask signal that skips the reading of photoelectrons. The register is updated for each frame. The local pixel driver, composed of a level

TABLE 1. TV for exposure times.

TV number	Exposure time	
	@1 frame CE	@16-frame CE
TV0	1/128 ms	1/128 ms
TV1	1/64 ms	1/64 ms
TV2	1/32 ms	1/32 ms
TV3	1/16 ms	1/16 ms
TV4	1/8 ms	1/8 ms
TV5	1/4 ms	1/4 ms
TV6	1/2 ms	1/2 ms
TV7	1 ms	1 ms
TV8	-	2 ms
TV9	-	4 ms
TV10	-	8 ms
TV11	-	16 ms

7 exposure stops
4 exposure stops
11 exposure stops

shift circuit and driver circuit, scans pixels in a predetermined row based on the decode signal of the integration time controller and SEL signal. Each readout unit is associated with 32 rows of pixels, corresponding to two exposure blocks per frame. In the "No Skip Mode," TX1 is sequentially controlled to cross two exposure blocks, and TX2 is independently controlled for each block according to the integration time. Short exposure times of one horizontal period or less can be achieved because the controls for TX1 and TX2 are independent. Moreover, "Skip Mode" can be realized by skipping the TX1 and TX2 operations using the mask signal. With the above configuration, the exposure time of each block was individually set by the integration time controller, and a rolling shutter reading of 16 × 32 pixels in each unit was performed simultaneously. Each block has an integration time controller that includes the selection of two modes. The two-dimensional matrix of the exposure time for each block (exposure-time table) was changed for each frame. Thus, with H264 × V264 resolution and 3-bit gradation, various exposure patterns can be created for each frame, resulting in HDR imaging. In addition, controlling exposure time over multiple frames makes the bit depth extend. Image acquisition can be performed by changing the exposure time for each block based on the exposure-time table, indicating the by time value (TV), which logarithmically defines the exposure time. The acquired image signals were multiplied and added according to TV. Consequently, images with a total of 11 exposure stops could be acquired at a high speed by setting the exposure time for seven exposure stops within one frame and four exposure stops for over a period of 16 frames.

Table 1 lists TVs for exposure times, which differ from the TV setting of a typical camera. In the case of one to seven exposure stops, which have an exposure time of TV0–TV7, the "No Skip Mode" is used. That is, the TX1 and TX2 operations never become skipped for every frame to control the exposure time within a frame. In the case of 11 exposure stops, the "No Skip Mode" is used for TV0–TV7, whereas the "Skip Mode" is used for TV8–TV11. In this mode, the TX1 and TX2 operations are skipped during the long exposure, crossing several frames.

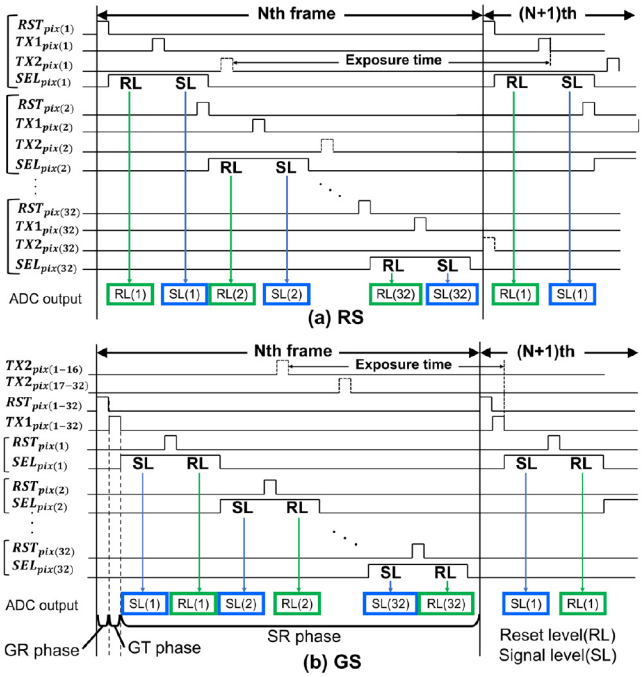


FIGURE 5. Timing chart of pixel read out.

E. TIMING CHART OF PIXEL READOUT

Fig. 5 shows timing charts of the pixel readout. The stacked sensor operates in a rolling shutter (RS) mode, which reads out 32 rows of pixel signals for each block sequentially row by row, or a global shutter (GS) mode, which uses the floating diffusion (FD) memory.

In the RS mode, the operation is the same as that of conventional sensors. As shown in Fig. 5(a), the reset level of the FD potential is first read out as a reference signal, followed by the readout of the PD signal level. Subsequently, the CDS operation is performed by subtracting the reset level from the signal level.

In contrast, the GS mode, the signal level is read first, then the reset level, as opposed to RS. The read noise is larger than that of the RS mode because CDS does not function effectively, that is, the threshold voltage (Vth) variation of the SF for each pixel can be canceled, but the kT/C noise of the RST cannot be removed. However, this configuration can reduce the pixel size without additional in-pixel memories.

The GS operation includes the global reset (GR), global transfer (GT), and sequential readout (SR) phases. In the GS mode, the read noise is increased because CDS does not function effectively. In the GR phase, RST is enabled in all pixels to reset FD. In the GT phase, TX1 is enabled in all pixels to transfer PD signals to FD. Subsequently, in the SR phase, after activating SEL for the first row and reading out the signal level to ADC, RST is enabled again to read out the reset level. This SR operation is repeated for 32 rows sequentially, and a digital CDS is performed in the peripheral circuitry. A dark current difference of FD occurs at the boundary between the 1st and 32nd rows, but it is small compared to the read noise and is at an insignificant level.

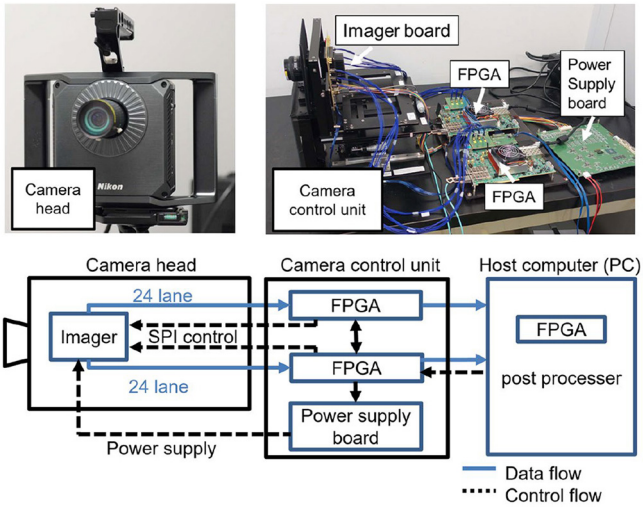


FIGURE 6. Camera system configuration.

In addition, the proposed CMOS image sensor (CIS) allows for frame skipping in both the RS and GS modes, which is effective for each exposure block, and allows for arbitrary exposure control in the readout operations, where the frame rate is not fixed.

III. CONFIGURATION OF THE EXPERIMENTAL SYSTEM

A. CAMERA SYSTEM

Fig. 6 shows the configuration of the system. This system comprises a camera head, camera control unit, and host personal computer (PC). The camera head has a lens unit and image sensor board. The camera control unit is composed of two FPGAs and a power supply board, and functions to control the image sensor, supply power, and receive image data. Moreover, PC controls the entire system and processes, and saves the images.

The image data from the sensor are input to each FPGA in the camera control unit via 24 channels at a time. FPGA adjusts the data rate and width, and transfers data to PC through an optical fiber. To calculate the exposure time of each block, we prototyped two systems based on the application. The first system calculates the exposure time in FPGAs of the camera control unit, assuming that the subject moves at a high speed. The exposure time is calculated based only on the light intensity of the acquired image. The calculation takes less than 1ms by using a pipeline operation without any frame memory, and the exposure time can be set differently for each frame, but since but the exposure time for 8 frames is transferred once every 8 frames, the update latency is 8 frames.

The second system calculates the exposure time in PC. More advanced and precise exposure control is possible considering the surrounding blocks or multiple frames. In addition, PC has another FPGA system inside, which enables the calculation and processing of both the software and FPGA (hardware) bases.

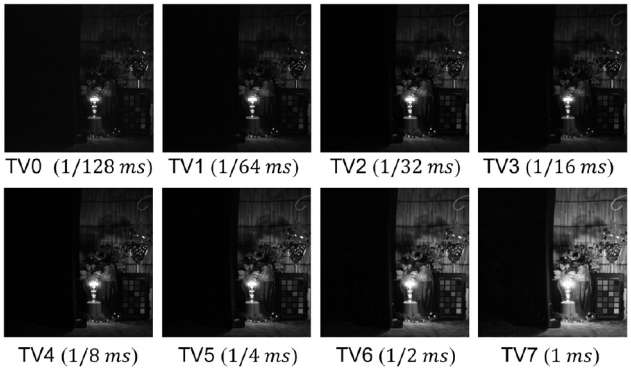


FIGURE 7. Diagram of the AE process.

B. IMAGE PROCESSING

1) AUTO EXPOSURE (AE) CONTROL

In order to obtain the optimum HDR image, it is necessary to perform AE to set the exposure time of each block before shooting. To capture the best HDR image, the exposure time of each block should be controlled to obtain the highest signal-to-noise ratio (SNR) in the block, without saturated conditions.

For further exploration of the starting point, the proper exposure-time table are explained using the photographs in Fig. 7. First, the exposure-time of all blocks is set to the same and pictures are captured by changing the exposure time from TV0-TV7. Next, the camera system determines on a block-by-block exposure time that is not saturated and maximizes the block's SNR. Finally, the exposure time of each block are recorded in the exposure-time table. If in an environment where the exposure state of the subject changes, many schemes can be used to efficiently obtain the optimum HDR image, such as by sequentially updating the exposure-time table during live view execution and a method to control exposure based on motion information in addition light intensities [6].

2) IMAGE PROCESSING FLOW

Fig. 8 shows the image signal processing (ISP) flowchart. The ISP flow contains the IF, DRAM control, ISP control, and several processing modules. These elements are controlled using a PC through the PCI buses.

First, the input signals are provided from the CIS output of HS-IF. These signals are rearranged in the raster format (H4224 × V4224, Bayer order), and stored in DRAM. Subsequently, several preconditioned corrections, such as DC offset and defect corrections, are applied to the image data.

Second, the flow is separated in the directions of HDR and block exposure control.

In the HDR flow, the frame addition is applied when frame rates are lower than 1000 fps. Moreover, CIS runs at 1000 fps and post-processing operates at 16-frames (62.5 fps) interval, as discussed in Section II-C. Subsequently, white-balancing and tone-mapping are applied. In the block exposure control flowchart, AE is applied and exposure-time table is provided to CIS through the exposure IF.

The input signal has a depth of 12 bits, HDR signal with single-frame CE has a depth of 19 bits, and HDR signal with a 16-frame CE has a depth of 23 bits. The final output signal after tone-mapping has a depth of 14 bits.

Consequently, all image data are processed in the raw data format (RAW); thus, no Bayer interpolation is applied until the final output. Furthermore, the colorized processes for input, intermediate, and output steps are performed in PC.

3) TONE CURVE

1) Photographic tone-reproduction operator: Based on the illumination or tonal range values, so-called the key values, of the scene, the logarithmic mean luminance is an approximation of the key values of the scene. Moreover, $\overline{L_\omega}$ is calculated as follows:

$$\overline{L_\omega} = \frac{1}{N} exp\left(\sum_{x,y} log(\delta + L_\omega(x, y))\right) \tag{2}$$

where $L_\omega(x, y)$ is the global luminance of pixel (x, y), N is the total number of pixels in the image, and δ is a small value used to avoid singularities that occur when black pixels in the image exist. When the scene has a normal key, the key is the mid-gray of the displayed image, or 0.18 on a scale of 0 to 1 [20]. In addition, it yields

$$L(x, y) = aL_\omega(x, y)/\overline{L_\omega} \tag{3}$$

where $L(x, y)$ is the scaled luminance and a is the is a key value that is chosen appropriate to the key of the scene.

2) Local sigmoidal tone-reproduction operator: In the case of digital images, applying the sigmoid operator with key values is possible for every pixel. The combination of simple tone-mapping and local sigmoidal operator yields the simple tone-mapping operation as simple tone mapping operation

$$L_d(x, y) = L(x, y)/\{1 + L(x, y)\} \tag{4}$$

and the Reinhard local operation as

$$L_d(x, y) = L(x, y)/\{1 + V(x, y)\} \tag{5}$$

where $L_d(x, y)$ is a display luminance and $V(x, y)$ is the average luminance of a local neighborhood of a certain size [20]. In the proposed system, operators for every block are applied as tone-mapping to the final RAW image data.

IV. SENSOR IMPLEMENTATION AND PERFORMANCE

A. IMPLEMENTATION

Fig. 9 shows a micrograph of the prototype image sensor. Top and Bottom chips were fabricated using a 65-nm process and stacked with each other. The die size is 18.87 $mm^{(H)}$ × 14.28 $mm^{(V)}$.

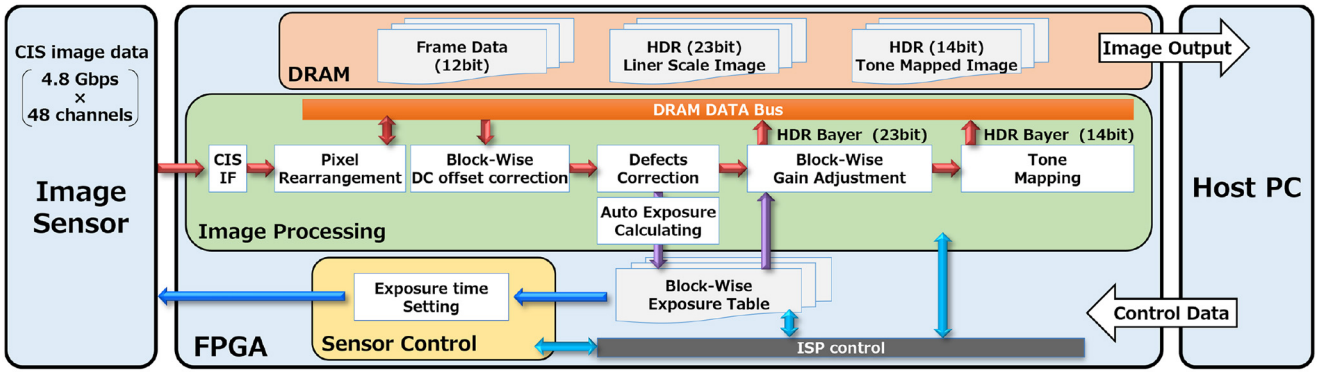


FIGURE 8. Image signal processing flow.

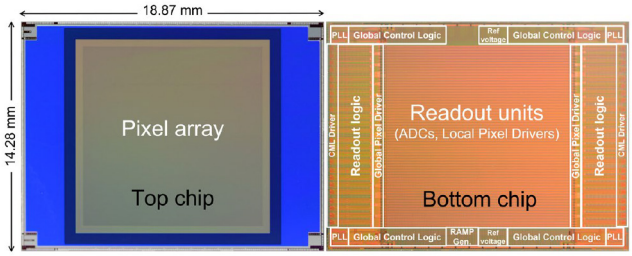


FIGURE 9. The die photographs.

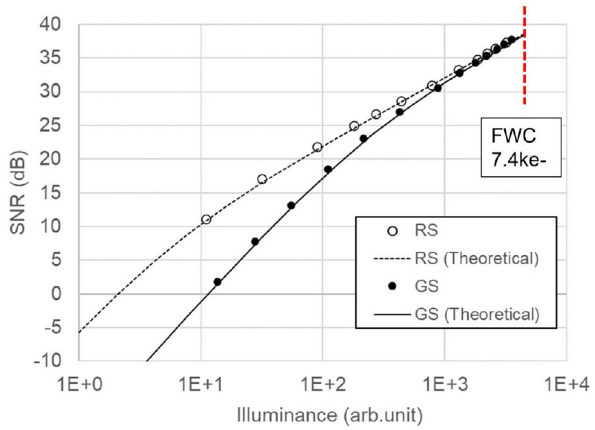


FIGURE 10. SNR in RS and GS.

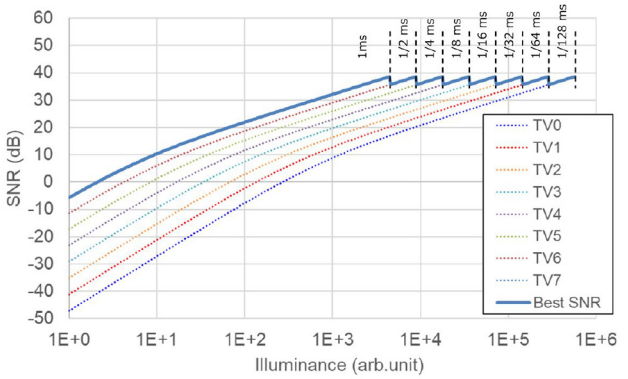


FIGURE 11. SNR with CE, single frame capture.

B. SENSOR PERFORMANCE

1) SNR OF RS AND GS MODES

The measured SNR in the RS and GS modes are shown in Fig. 10. The exposure time was 1 ms for all blocks. The right end of the theoretical curve shows the full well capacity (FWC). The theoretical curve for SNR is expressed as follows:

$$SNR = 20\log\left(Q_t/\sqrt{N_{shot}^2 + N_{dark}^2 + N_r^2}\right)$$
$$\because N_{dark} \ll N_r$$
$$\cong 20\log\left(Q_t/\sqrt{N_{shot}^2 + N_r^2}\right) \quad (6)$$

where Q_t, N_{shot}, N_{dark} and N_r are the number of signal electrons, the shot noise expressed as $\sqrt{Q_t}$, the dark current noise, and read noise. For an exposure time of 1 ms, N_{dark} is negligible because it is sufficiently smaller than N_r. The measured N_r in the RS and GS modes is 2.9 e^-_{rms} and 18.4 e^-_{rms}, respectively. The number of signal electrons under the same illumination has the same number of electrons in both modes. In sufficiently bright conditions, N_{shot} is the dominant factor in the total noise; therefore, SNRs in the RS and GS modes are approximately identical. Moreover, in dark conditions, N_r is the dominant factor in the total noise; thus, SNR in the RS mode is higher than that of the GS mode.

2) SNR WITH SINGLE-FRAME CE

Fig. 11 shows SNR with block-wise CE which demonstrates the calculation results for seven exposure stops at 1000 fps in the "No Skip Mode" based on the measured SNR in the RS mode, as shown in Fig. 10. To obtain a high SNR, it is desirable to use the longest exposure time. However, if the signal saturates, the correct output cannot be obtained. Signal saturation occurs in PDs, signal detectors, and ADCs, all of which can be prevented by reducing the exposure time. Having seven exposure stops with single frame expands the saturation signal level 2^7 times equivalently. Therefore, an HDR photography is possible by an appropriate adjustment of the exposure time. SNR with single-frame CE (SNR_{CE})

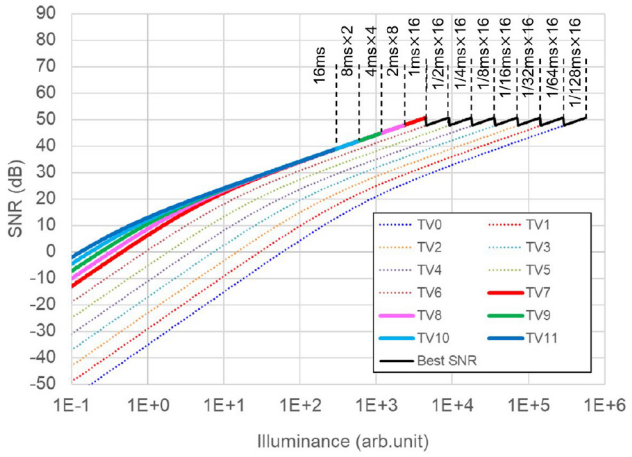


FIGURE 12. SNR with CE, 16 frames capture.

can be calculated as follows:

$$SNR_{CE} \cong 20\ \log\left(\alpha Q_t/\sqrt{(\alpha N_{shot})^2 + (\alpha N_r)^2}\right)$$
$$\alpha = 2^{(7-N)}$$
$$= 20\ \log\left(Q_t/\sqrt{N_{shot}^2 + N_r^2}\right) \quad (7)$$

where α is the multiplication of the exposure time normalized by 1 ms to a power of 2 and N is the TV number. The acquired image is reconstructed by multiplying the gain according to the TV number, but since both the signal and noise are multiplied by the gain, the SNR can be expressed in the same as in (6). The graph of "Best SNR" shows the best case using the highest SNR condition. Because FWC does not increase, SNR becomes a jagged graph, as shown in Fig. 11.

3) SNR WITH A 16-FRAME CE

Fig. 12 shows SNR for a case with 11 exposure stops. Exposure time is in the range of TV0–TV11. The image signal is captured at a readout operation of 1000 fps and obtained using 16 frames in a combination of the "Skip Mode" and "No Skip Mode." For seven exposure stops from TV0 to TV7, the image signals are obtained in every frame of 1000 fps using the "No Skip mode." In the TV8 condition, signal electrons are integrated during two frames in the "Skip Mode." This implies that the exposure time was 2 ms, skipping a single frame. At TV9, signal electrons are integrated during four frames, which have an exposure time of 4 ms, skipping three frames. Similarly, by increasing the number of skipping frames to TV11 (15 frames), signal electrons are integrated, and finally all image signals are obtained at the end of the 16^{th} frame. In the TV11 condition, the total number of signal electrons increases by a factor of 16, and the noise component increases by a factor of 4 ($=\sqrt{16}$) for the same illuminance. Therefore, SNR expands by approximately 12 dB compared to the single-frame CE

Additionally, when the exposure time is shorter than 16 ms, the signals can be read multiple times, and each readout signal can be integrated. Therefore, the total exposure time for TV7–TV11 is 16 ms. TV11 uses the frame skipping to store 16-ms worth of charge in PD and reads it out only once. In contrast, TV7 reads out 16 times every 1 ms and accumulates them. Moreover, performance tradeoffs between a longer integration time of more than 1 ms (a single-frame capture) and a higher number of accumulations (multi-frame captures) at the same exposure time exist. Because the exposure time at TV11 is 16 times longer than TV7 (1 ms), TV11 has the same number of signal electrons with 1/16 of the illumination as TV7, yielding the same SNR. Therefore, it is more resistant to darkness.

However, the signal saturation is traded off because TV11 accumulates the signal for 16 ms and reads out only once; thus, the signal saturates when exceeds FWC. Therefore, TV11 can be used only under low-illumination conditions, where it does not saturate. In the case of adding 16 frames at TV7, the readout and accumulation are performed 16 times for 1 ms each; thus, the signal can be read out up to 16 times higher illuminance than that of TV11 without saturation. The maximum total number of signal electrons is 16 × FWC. However, because 16 readouts are performed, the readout noise increases by a factor of 4 ($=\sqrt{16}$). Therefore, SNR in dark areas is worse than that in TV11.

The characteristics of TV8–TV10 is between those of TV7 and TV11. The SNR with a 16-frame CE (SNR_{16-CE}) equation can be expressed as follows:

$$SNR_{16-CE} = 20\ \log\left(Q_{t16}/\sqrt{Q_{t16} + N_{dark16}^2 + nN_r^2}\right)$$
$$\cong 20\ \log\left(Q_{t16}/\sqrt{Q_{t16} + nN_r^2}\right) \quad (8)$$

where Q_{t16} is the number of signal electrons accumulated during the 16-frame period, N_{dark16} is the number of dark-noise electrons during the 16-frame period, and n is the number of readouts during the 16-frame period, that is, 16 readouts without frame skips, and 1, 2, 4, or 8 readouts with frame skips. For an exposure time of 16ms, the dark current noise is less than 1/5 of the readout noise and is a small value, so it can be ignored. The "best SNR" graph shows the best case using the highest SNR condition from TV6 to TV0.

4) BLOCK TO BLOCK NOISE DIFFERENCE

Fig. 13 shows the block steps at the adjacent block boundary when in a dark scene a point light source exists. To avoid the saturation of blocks with the point light source, they should be shot with a short exposure (e.g., 1/128 ms; TV0). However, its neighbor block, has no high-intensity objects. To maximize the SNR of the neighbor block, it should be captured with a long exposure (e.g., 1 ms; TV7). The boundary area between light source blocks and the neighbor blocks is continuous illuminance. However, as is clear from comparing the graphs of TV0 and TV7 in Fig. 10, a significant change of TV causes a large SNR step at the

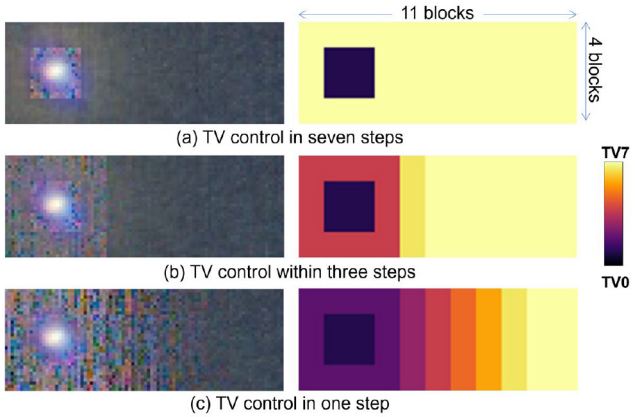


FIGURE 13. Steps at the adjacent block boundary.

TABLE 2. Summary of sensor characteristics.

Item	Data		
Process	top: 65nm, bottom: 65nm		
Chip size	18.87 (H) mm × 14.28 (V) mm		
Number of pixels	4224 (H) × 4224 (V)		
Pixel pitch	2.7 μm		
Output interface	4.8 Gbps/channel × 48 channels		
FWC	7.4 ke$^-$		
ADC resolution	12 bit		
Sensitivity	20.7 ke$^-$/lux·s		
Conversion gain	161 μV/e$^-$		
Random noise	2.9 e^-_{rms} @RS		
	18.4 e^-_{rms} @GS		
Frame rate	1000 fps		
Dynamic range	68dB w/o CE	110 dB 1 frame CE	134 dB 16 frames CE
Power consumption	7.4 W		

block boundaries. Therefore, it is preferable to control the TV settings appropriately in adjacent blocks. In the prototype camera system, the exposure time in adjacent blocks is controlled in one or three steps. The number of steps required to limit the exposure time in adjacent blocks can be changed by parameter setting. The image quality near point sources gives priority to preventing block steps rather than improving SNR.

Table 2 summarizes the performance of the prototype image sensor. The DR without CE was 68 dB, which can be expanded by 110 dB (2^7 times or 42 dB) with a single-frame CE of 1000 fps, and by 132 dB (2^{11} times or 66 dB) with 16-frames CE. The supply voltages are 3.3V for analog domain and 1.8V and 1.25V for digital domain and the power consumption at 1000 fps is 7.4 W.

V. EXPERIMENTAL RESULTS FOR HDR IMAGING

A. BLOCK EXPOSURE

Fig. 15 shows the experimental results of CE. To generate an exposure-time table, the original image was first divided into 264 × 264 blocks, and the average value of each block was converted into 3-bit gradation data. Subsequently, the exposure time was set from 1/128 to 1 ms, corresponding to the 3-bit data. A photography was performed on

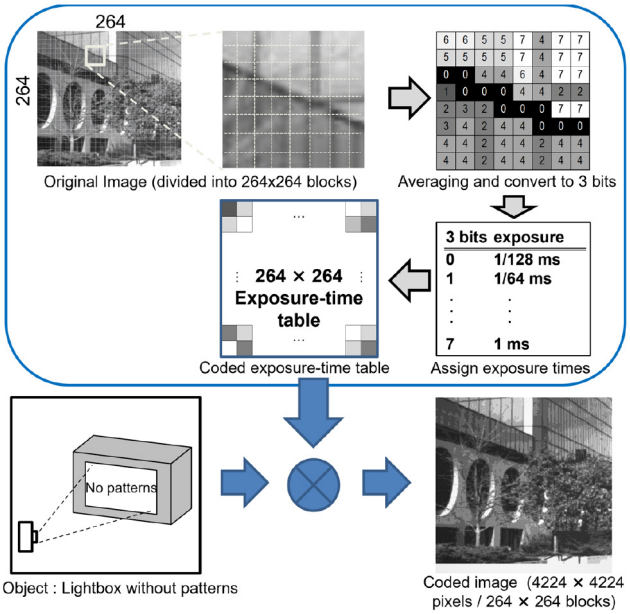


FIGURE 14. Experimental results of CE.

a light box without patterns under uniform illumination using an exposure-time table. The block-exposure can produce a coded image in a single shot. Furthermore, setting the exposure time frame by frame or across frames yields various coded patterns.

B. HDR IMAGING

1) HDR IMAGE SYNTHESIZING BY EXPOSURE BRACKETING

Fig. 14(a) presents images obtained by changing the exposure time without CE. This approach is known as exposure bracketing or multiple exposures, which is used for the synthesis of HDR images [21]. Several images with different exposure times were prepared. Moreover, DR of each image is relatively small (68 dB in the developed sensor); therefore, blackouts and whiteouts occur. Therefore, many images are required; thus, a long time to capture images is needed. Subsequently, we scanned all image data at each exposure and found and recorded signals of the highest SNR regions, excluding the saturated regions. When saturated regions were obtained, the data were left as is and recorded even though the minimum exposure time was applied in the camera setting. Finally, HDR was generated while collecting the regions with the highest SNR and synthesizing them into one image.

2) HDR IMAGE BY SINGLE FRAME CE

By comparison, as shown in Fig. 14(b), an HDR image is acquired and calculated with CE using the proposed sensor and post-processing system. In the single frame mode, the reconstruction is shown in the following equation,

$$E'(i,j) = E(i,j) \times \alpha$$
$$\alpha = 2^{(7-N)} \quad (9)$$

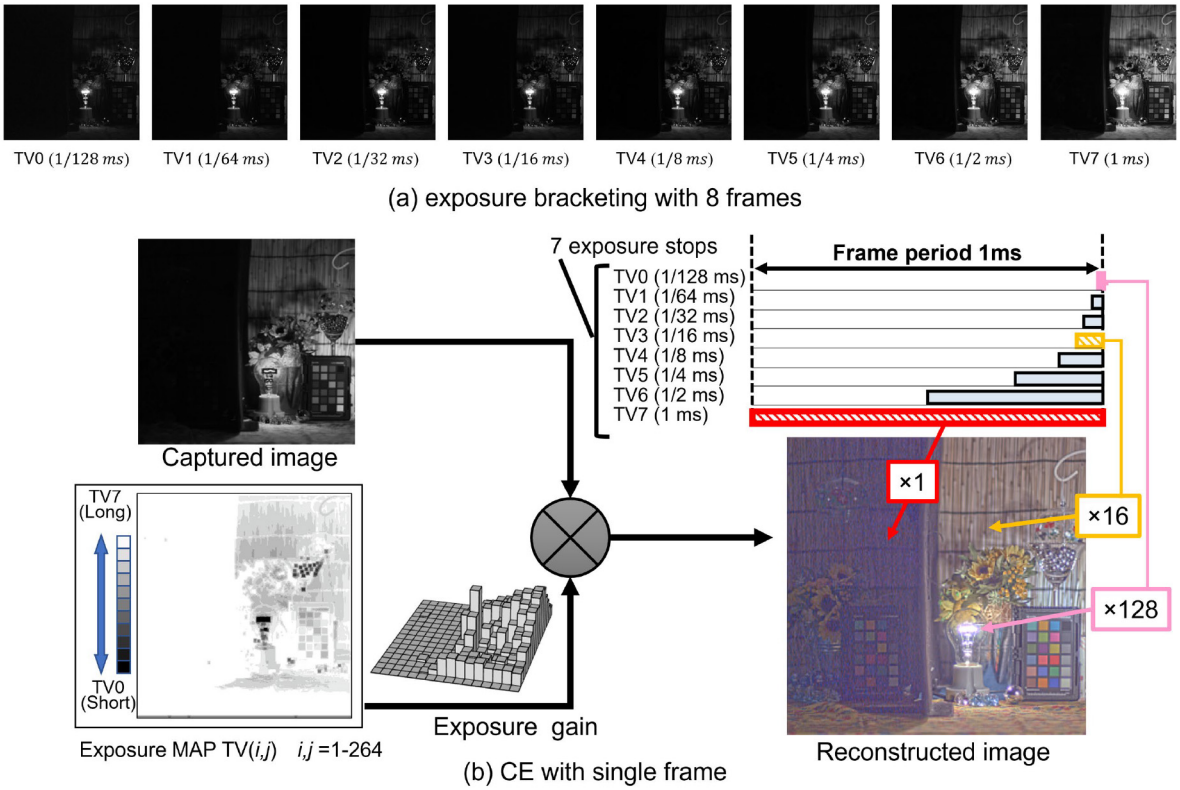


FIGURE 15. Experimental results of HDR with single frames using CE in RS mode.

where $E(i, j)$ is a data matrix for each block of captured image, i and j are array numbers of exposure blocks, $E'(i, j)$ is a data matrix of reconstructed image and N is the TV number of each block. The calculation is applying the gain table to the captured image data yields a high DR image. The gain table is simply the inverse number of exposure times (TV number). By using block-wise CE, a 110 dB image can be acquired in one shot (1 ms).

3) HDR IMAGE BY 16-FRAME CE

In the 16-frame mode, as shown in Fig. 16, the sensor was driven at 1000 fps, and the post-processing system operation interval was 16 frames, including the skipping frames. Image acquisition was performed by changing the exposure time for each block based on the exposure-time table, denoted as TV, which logarithmically defines the exposure time. The reconstruction performs the following equation,

$$E'(i,j)_{(TV0-6)} = \sum_{frame=1}^{16} E(i,j) \times \alpha$$
$$E'(i,j)_{(TV7-11)} = \sum_{frame=1}^{M} E(i,j)$$
$$M = 2^{11-N} \tag{10}$$

As shown in (10), multiplication and addition processing of the acquired image according to the TV. The frame skipping scheme is achieved using the mask bit, as discussed in Section II-C. For example, region A in the image data is exposed in t = 16 ms by skipping 15 frames, and region B is exposed in t = 8 ms and added to a single frame by skipping seven frames twice. Similarly, decreasing the exposure time and adding several frames into a single frame is performed until an exposure time of 1 ms. When the exposure time is shorter than 1 ms, the frames are simply applied gains and added 16 times. Therefore, images with a total of 11 exposure stops can be acquired at 62.5 fps by setting the exposure time for seven exposure stops within one frame and four exposure stops over a period of 16 frames.

C. ADAPTIVITY FOR MOVING OBJECT

Fig. 17 shows the responsiveness of the exposure control to a moving object in RS mode. The sensor was operated at 1000 fps, while the object was horizontally moving on the screen. Without exposure control shown in Fig. 17(a), the body text becomes oversaturated when the object dashes from dark to light areas at frame #345 relative to the initial state at frame #1. With the dynamic exposure control in every eight frames shown in the Fig. 17(b), updating the exposure-time table every eight frames suppresses the oversaturation of characters on the body and realizes an HDR image.

On the other hand, artifacts due to the latency of exposure update can be confirmed at boundaries of movement such as near the front wheel. The major factor of the latency is SPI communication between FPGA and the image sensor. So, speeding up the SPI or updating the exposure table only

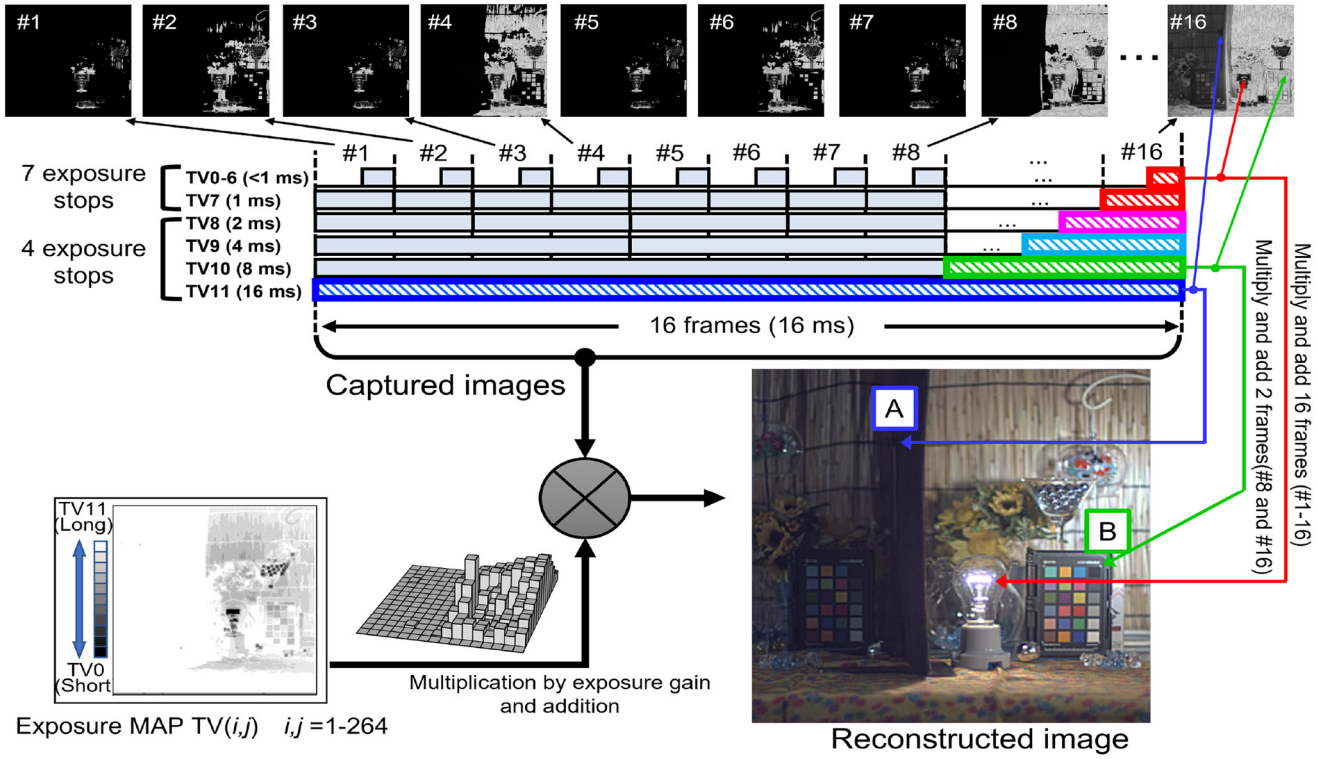


FIGURE 16. Experimental results of HDR with 16 frames using CE in RS mode.

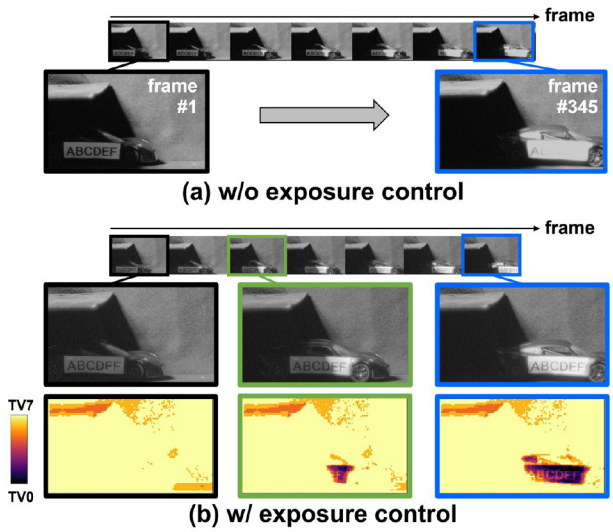


FIGURE 17. Experimental results for moving objects.

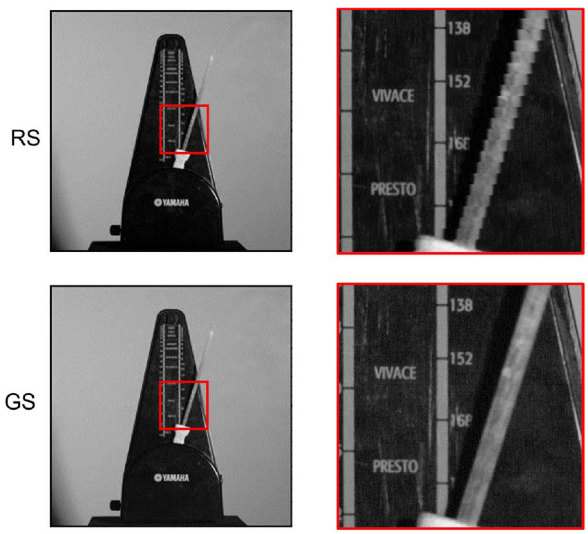


FIGURE 18. Comparison of RS and GS.

in motion areas reduces the artifacts, and prediction using motion vectors is also useful.

Fig. 18 shows the experimental results of shooting moving objects in RS and GS modes. In RS mode, high-speed shooting at 1000 fps suppresses the shutter distortion across the whole image that occurs with conventional sensors. However, the partial shutter distortion occurs every 32 rows due to the block-based readout operation. In GS mode, the above distortion is eliminated.

D. PERFORMANCE COMPARISON WITH EXISTING SENSORS

Table 3 presents a comparison of the performance of the proposed sensor and that of other sensors presented in the existing studies. First, our developed sensor achieved a high-speed readout with a resolution of 17 Mpixels. Energy efficiency is superior in terms of figure of merit (FoM1). In addition, this study demonstrated an HDR with a small pixel of 2.7-μm pitch.

TABLE 3. Comparison with existing sensors.

	This work (1 frame RS)	This work (16 frame RS)	This work (1 frame GS)	[5]	[6]	[8]	[9]	[17]
Process	Stacked BSI Top: 65nm Bottom: 65nm			Stacked BSI Top: 90nm/65nm Bottom: 40nm	Stacked BSI Top: 90nm Bottom: 55nm	FSI 0.18μm	0.11μm	Stacked BSI Top: 45nm Bottom: 65nm
HDR technology	Integration time controllable for each block			Sub-pixel	-	Integration time controllable for each block	Integration time controllable for each pixel	Time to saturation and dual CG for each pixel
# of pixels (Mpix)	17.8			5.7	4.1	0.4	0.05	0.25
Pixel pitch (μm)	2.7			3.0	4.8	5.0	11.2	4.6
Bit depth (bit)	12			12	-	-	-	10
Frame rate (fps)	1000			30	630	-	25	30
Conversion gain (μV/e-)	161			6.7 (Low), 197 (High)	65	-	-	7 (Low), 170 (High)
Sensitivity (ke-/lux · s)	20.7			38.0	28.4	-	-	-
FWC (ke-)	7.4			165.8	-	-	-	9000*
Power (mW)	7400			-	185	-	34.4	5.75
Shutter mode	RS		GS	RS	GS	RS	GS	GS
Random noise (e-)	2.9 (0dB)		18.4 (0dB)	0.6	4.2 (24dB)	-	-	4.2(0dB)
DR (dB)	110 †	134 †	94 †	132	-	120	-	127
FoM1 (e- · pJ/step)	0.29	4.70	1.86	-	-	-	-	3.00
FoM2 (e- · pJ/DRU)	0.0037	0.0037	-					0.0014

$FoM1 = (power \times noise)/\{\#\ of\ \ pixels \times frame\ rate\ \times 2\text{^}(bit\ depth)\}$

$FoM2 = (power \times noise)/(\#\ of\ \ pixels \times frame\ rate\ \times DRU);\ DRU = \{(saturation/gain)/noise\}$

*:Equivalent FWC with photo response plot

†: $DR@1\ frame\ CE = 20 \times \log(FWC \times 2^7/noise)$

†: $DR@16\ frame\ CE = 20 \times \log(FWC \times 2^{11}/noise)$

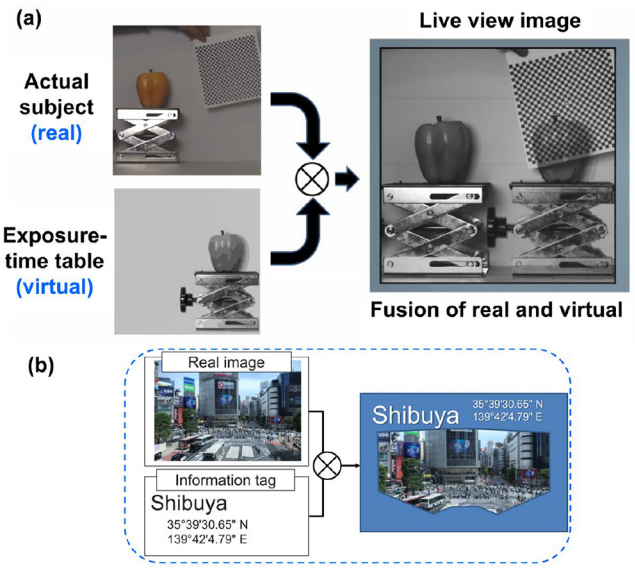


FIGURE 19. (a) Experimental results of the overlay information on the real image and (b) an application example of CE.

FIGURE 20. Application example for security.

E. MORE APPLICATIONS USING CODED EXPOSURE

Fig. 19 shows an application that overlays the information on a real image. Encoding the exposure time for each region implies that an arbitrary virtual pattern can be generated in an image. The experimental results are shown in Fig. 19(a). The photograph on the upper left shows the actual subject. The lower-left figure presents an image of the exposure map, where a pseudo image is formed by expressing eight different exposure settings as gradations. The dark and bright areas indicate a short and long exposure, respectively. The results obtained using this exposure map are shown on the right side. It can be observed that a virtual image was embedded in the subject in the real space. In addition, this exposure-time table was created based on an actual subject; thus, it can be used as an image-recording function.

Fig. 19(b) shows an application where it is possible to embed a tag information or head-mounted display for virtual or augmented reality [22], [23]. Overlaying them on the image sensor reduces the processing power of the subsequent system and improves the latency.

Fig. 20 shows another application example for security or privacy protection. The exposure-time table is used as a decryption key to reconstruct the image correctly. Left images were captured by using CE which the exposure-time are mapped in random manner spatially varied from TV0 to TV7. These images are very noisy and too difficult to recognize the details of the characters or of the sunflowers.

This is caused by the exposure time difference in every block. So, improving the random noise by a simple global control of gain or tone is difficult.

Right images are reconstructed by using the exposure-time table. Apparently, the image is cleaned, and the characters are easily to recognize. This is done by adjusting the exposure gain each block to remove the exposure time difference. This example is just showing a spatially random pattern, but in more advanced, the combination of the spatial and the temporal patterns is possible and can be more reliable in the decryption key.

VI. CONCLUSION

In this study, we developed a CMOS image sensor with a stacked structure that operated at 1000 fps, while yielding a high resolution of 17 Mpixel and a small pixel of 2.7-μm pitch. Using the block-wise CE function, 110 dB DR were achieved at 1000 fps in a single frame CE. Moreover, a DR can be expanded to134 dB with16-frame CE. This vision system can be applied to various computational imaging techniques using the CE function in addition to high-dynamic-range imaging in scenes where dark and bright areas of the subject are mixed in the frame.

ACKNOWLEDGMENT

The authors would like to thank all engineers at the Nikon Corporation Advanced Technology Research & Development Division, Opto Device Development Center, and Mathematical Sciences Research Laboratory for their support in this work.

REFERENCES

[1] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: Motion deblurring using fluttered shutter," in *Proc. ACM SIGGRAPH*, 2006, pp. 795–804.

[2] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. Nayar, "Video from a single coded exposure photograph using a learned over-complete dictionary," in *Proc. IEEE ICCV*, 2011, pp. 287–294.

[3] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proc. ACM SIGGAPH Conf. Comput. Graph. Interact. Techn.*, 1995, pp. 39–46.

[4] K. Miyauchi, S. Okura, K. Mori, I. Takayanagi, J. Nakamura, and S. Sugawa, "A high optical performance 2.8μm BSI LOFIC pixel with 120ke$^-$ FWC and 160 μV/e conversion gain," in *Proc. Int. Image Sensor Workshop*, 2019, pp. 246–249.

[5] Y. Sakano et al., "A 132dB single-exposure-dynamic-range CMOS image sensor with high temperature tolerance," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 106–108.

[6] T. Takahashi et al., "A 4.1Mpix 280fps stacked CMOS image sensor with array-parallel ADC architecture for region control," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2017, pp. 244–245.

[7] S. Velichko et al., "140 dB dynamic range sub-electron noise floor image sensor," in *Proc. Int. Image Sensor Workshop*, 2017, pp. 294–297.

[8] A. Peizerat et al., "A 120dB DR and 5μm pixel pitch imager based on local integration time adaptation," in *Proc. Int. Image Sensor Workshop*, 2015, pp. 385–388.

[9] N. Sarhangnejad et al., "Dual-tap pipelined-code-memory coded-exposure-pixel CMOS image sensor for multi-exposure single-frame computational imaging," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 102–104.

[10] J. Zhang, T. Xiong, T. Tran, S. Chin, and R. Etienne-Cummings, "Compact all-CMOS spatiotemporal compressive sensing video camera with pixel-wise coded exposure," *Opt. Exp.*, vol. 24, no. 8, pp. 9013–9024, 2016.

[11] Y. Luo and S. Mirabbasi, "A 30-fps 192 × 192 CMOS image sensor with per-frame spatial-temporal coded exposure for compressive focal-stack depth sensing," *IEEE J. Solid-State Circuits*, vol. 57, no. 6, pp. 1661–1672, Jun. 2022.

[12] M. Yoshida, T. Sonoda, H. Nagahara, K. Endo, Y. Sugiyama, and R.-I. Taniguchi, "High-speed imaging using CMOS image sensor with quasi pixel-wise exposure," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, Evanston, IL, USA, May 2016, pp. 1–11.

[13] J. Zhang et al., "A closed-loop, all-electronic pixel-wise adaptive imaging system for high dynamic range videography," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 6, pp. 1803–1814, Jun. 2020.

[14] H. M. So, J. N. P. Martel, P. Dudek, and G. Wetzstein, "MantissaCam: Learning snapshot high-dynamic-range imaging with perceptually-based in-pixel irradiance encoding," 2021, *arXiv:2112.05221*.

[15] S. Nayar and V. Branzoi, "Adaptive dynamic range imaging: Optical control of pixel exposures over space and time," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1168–1175.

[16] H. Ke et al., "Extending image sensor dynamic range by scene-aware pixelwise-adaptive coded exposure," in *Proc. Int. Image Sensor Workshop*, 2019, pp. 111–114.

[17] R. Ikeno et al., "A 4.6-um, 127-dB dynamic range, ultra-low power stacked digital pixel sensor with overlapped triple quantization," *IEEE Trans. Electron Devices*, vol. 69, no. 6, pp. 2943–2950, Jun. 2022,

[18] T. Hirata, H. Murata, H. Matsuda, Y. Tezuka, and S. Tsunai, "A 1-inch 17Mpixel 1000fps block-controlled coded-exposure back-illuminated stacked CMOS image sensor for computational imaging and adaptive dynamic range control," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 120–121.

[19] T. Hirata et al., "A 1-inch 17Mpixel 1000fps block-controlled coded-exposure back-illuminated stacked CMOS image sensor for computational imaging and adaptive dynamic range control," Nikon Res., Tokyo, Japan, Rep., vol. 4, Sep. 2022.

[20] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[21] E. Reinhard et al., "HDR image capture," in *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, 2nd ed. Burlington, MA, USA, Morgan Kaufmann, 2010, pp. 145–204

[22] J. Wang et al., "Augmented reality navigation with automatic marker-free image registration using 3-d image overlay for dental surgery," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1295–1304, Apr. 2014.

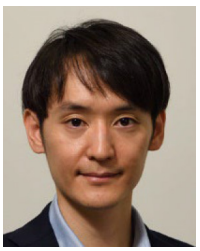
[23] H. Hile and G. Borriello, "Positioning and orientation in indoor environments using camera phones," *IEEE Comput. Graph. Appl.*, vol. 28. no. 4, pp. 32–39, Jul./Aug. 2008.

TOMOKI HIRATA received the B.E. and M.E. degrees in electric engineering from Hiroshima University, Hiroshima, Japan, in 2005 and 2007, respectively. He joined Nikon Corporation in 2007, and is involved in research and development of image sensors, especially in design of analog circuits.

HIRONOBU MURATA started research and development work with Texas Instruments Japan in 1991, developing and designing CCD/CIS/AFE-IC. From 2002 to 2008, he was developing custom CCD/CIS with Eastman Kodak Company. Since 2008, he has been engaged in camera design and imaging system research and development with Nikon Corporation, Japan.

TAKU ARII joined Nikon Corporation, Japan, in 2007, and is involved in the design of CMOS image sensors.

HIDEAKI MATSUDA joined Nikon Corporation, Japan, in 1993, and is involved in the development of CMOS image sensors.



HAJIME YONEMOCHI received the B.S. and M.S. degrees in science and engineering from Kanazawa University, Ishikawa, Japan, in 2012 and 2014, respectively.

In 2014, he joined Nikon Corporation, Japan, and is involved in the development of CMOS image sensors.

YOJIRO TEZUKA received the B.S. and first M.S. degrees in material science from the University of Tokyo, Tokyo, Japan, in 2000 and 2002, respectively, and the second M.S. degree in management of technology from the Tokyo University of Science, Tokyo, Japan, in 2022. He joined Nikon Corporation in 2002, and is involved in research and development of CMOS imagers, especially in designs of analog circuits.

SHIRO TSUNAI received the B.E. and M.E. degrees in electrical engineering from the Musashi Institute of Technology, Tokyo, Japan, in 1988 and 1990, respectively. From 1990 to 2008, he was with NEC Corporation, Japan, where he was involved in development of architectures, circuits and devices for image sensors. In 2008, he joined Nikon Corporation, Japan, where he was involved in development of digital still camera. He is currently in charge of Research and Development of CMOS image sensors.