PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Sound capture from rolling-shuttered visual camera based on edge detection

Koichi TERANO[1]; Hiroki SHINDO[2];

Kenta IWAI[3]; Takahiro FUKUMORI[4]; Takanobu NISHIURA[5]

[1, 2] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

[3, 4, 5] College of Information Science and Engineering, Ritsumeikan University, Japan

ABSTRACT

Recently, several new studies for recording sound have been conducted on extracting sounds from images of an object surface vibrated by sound waves. This method can capture the sound by a high-speed camera instead of an air conduction microphone. However, this method is unpractical due to higher cost of the high-speed camera. In this paper, we propose a new sound capture method with lower cost using a rolling-shuttered visual camera. This camera uses a CMOS (complementary metal-oxide-semiconductor) image sensor and sequentially writes an image from the top row to the bottom row. Therefore, when a moving object is photographed, a rolling-shutter distortion occurs due to the different writing time for each row. The proposed method uses the edge detection to emphasize the edge of the rolling-shutter distortion caused by photographing the vibrating object. Then, the proposed method resamples the edge as the amplitude of the sound wave. As a result of the sound capturing experiment, we confirmed that the proposed method can capture pure tone sound below 1,000 Hz from the image photographed by the CMOS image sensor.

Keywords: Sound capture, Rolling-shuttered visual camera, CMOS image sensor, Edge detection

1. INTRODUCTION

A typical air conduction microphone captures not only the target sound but also the noise emitted from the sound sources surrounding of the microphone. Recently, for the purpose of only target sound capture, several sound capture methods by measuring the vibration of objects affected by sound waves have been proposed [1, 2, 3]. In [1], a laser Doppler vibrometer is used and captures sound by irradiating a vibrating object with laser light and measuring difference between the reflected light and the reference light. In [2, 3], a high-speed camera is used and directly captures sound by photographing the vibration of the object and measuring the displacement of surface. In these methods, it is possible to capture only the target sound by measuring the vibration of the object near the target sound source. However, these methods are unpractical due to high cost of the laser Doppler vibrometer or that of the high-speed camera.

We propose a new sound capture method from the image of the vibrating object affected by sound waves with a rolling-shuttered visual camera. This camera is widely used in digital cameras and smartphone cameras [4]. Thus, the cost of the proposed method is lower than that of the conventional methods. This camera uses a small image sensor called complementary metal-oxide-semiconductor (CMOS) image sensor that senses light and converts it into an electrical signal. The CMOS image sensor sequentially writes an image from the top row to the bottom row. Therefore, when photographing a moving object, rolling-shutter distortion occurs in the image due to the difference in photographing time between each row [5]. In this paper, we use the rolling-shuttered visual camera to photograph the surface of a vibrating object. The proposed method uses the edge detection to emphasize the edge of rolling-shutter distortion caused by sound waves, and measures displacement

[1] is0267rs@ed.ritsumei.ac.jp
[2] is0261xf@ed.ritsumei.ac.jp
[3] iwai18sp@fc.ritsumei.ac.jp
[4] fukumori@fc.ritsumei.ac.jp
[5] nishiura@is.ritsumei.ac.jp

of vibration to capture the sound affecting the object. We conducted an experiment to confirm the performance of the proposed method.

2. Image capture with CMOS image sensor

In this paper, we use the rolling-shuttered visual camera to photograph the surface of a vibrating object affected by sound waves. This camera is widely used in digital cameras that built into smartphone because of its small size and low power consumption [4]. It employs a CMOS image sensor as the imaging device. When the moving object is photographed by this sensor, rolling-shutter distortion occurs due to the difference in photographing time between each row. The principle of the CMOS image sensor and rolling-shutter distortion are shown in Fig. 1. Photographing with CMOS image sensor is performed by energizing and exposing a photodiode disposed in a pixel and converting received light into electrical signal. In order to reduce the power consumption at photographing, the CMOS image sensor performs exposure by energizing the photodiodes sequentially from the top row to the bottom row. Therefore, when photographing a moving object, the position of the object changes during exposure from row to row. This positional deviation is called rolling-shutter distortion [5]. Sound waves propagate to the environment as air pressure fluctuations. Hence, the surface of the object near the sound source vibrates due to affection of the sound wave. The exposure sampling frequency of the CMOS image sensor is sufficiently higher for human audible sound. As a result, the rolling-shuttered visual camera can photograph surface vibration of the object affected by sound waves as rolling-shutter distortion. In other words, we can capture the sound affecting objects by detecting the rolling-shutter distortion caused by sound waves on the image of vibrating object.

3. Sound capture based on edge detection from the image

In this paper, we propose a new sound capture method by measuring the displacement of the vibrating object affected by sound waves based on the edge detection for the rolling-shutter distortion. For simply explanation, we assume the situation of photographing straight lines on a paper in this section. The overview of the proposed method is shown in Fig. 2. The photographed image is stored as a grayscale image source of a matrix having only light intensity values 0-255 for each pixel. At first, the proposed method detects the edge of the grayscale image using Canny operator [6] which is one of the most widely used because of its robustness to various image size [7, 8]. The edge detection using Canny operator has four steps as follows.

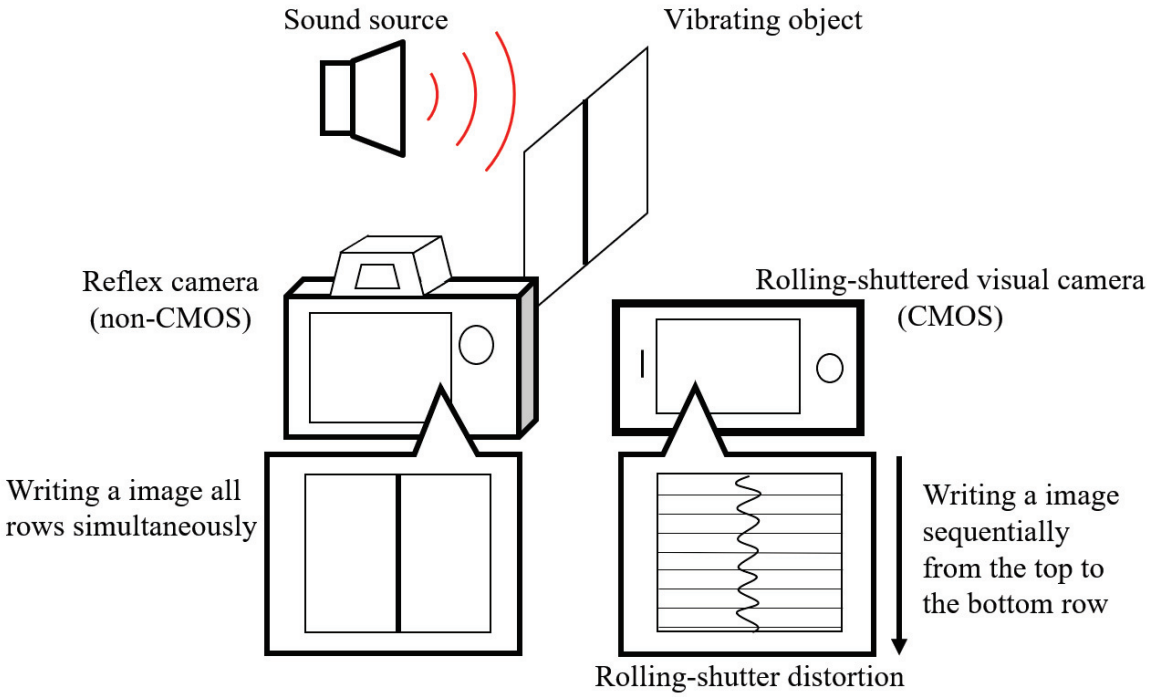


Figure 1 – Principle of CMOS image sensor and rolling-shutter distortion.

Step 1: Low pass filtering with Gaussian filter

Gaussian smoothing filter is used to remove impulse noise on the image. In this paper, we use the 3×3 neighborhood kernel which is one of the most widely used for Gaussian filter. The 3×3 neighborhood image source $\boldsymbol{V}_{x,y}$ and the kernel of Gaussian filter $\mathbf{G}$ are respectively shown in Eqs. (1) and (2). The smoothed image $s_{x,y}$ by Gaussian filter is given by Eq. (3).

$$\boldsymbol{V}_{x,y} = \begin{bmatrix} v_{x-1,y-1} & v_{x,y-1} & v_{x+1,y-1} \\ v_{x-1,y} & v_{x,y} & v_{x+1,y} \\ v_{x-1,y+1} & v_{x,y+1} & v_{x+1,y+1} \end{bmatrix}, \tag{1}$$

$$\mathbf{G} = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \tag{2}$$

$$s_{x,y} = \boldsymbol{V}_{x,y} * \mathbf{G}, \tag{3}$$

where $v_{x,y}$ is a pixel value of the input image at coordinate (x, y) and symbol '$*$' is a convolution operator.

Step 2: Image gradient calculation with Sobel filter.

Sobel filter is used to calculate the gradient and gradient direction of each pixel. Each directional kernel of Sobel filter $\mathbf{G}_\mathrm{x}, \mathbf{G}_\mathrm{y}$ is shown in Eq. (4).

$$\mathbf{G}_\mathrm{x} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \mathbf{G}_\mathrm{y} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{4}$$

The gradient $p_{x,y}$ and the gradient direction $\theta_{x,y}$ of each pixel are given by Eqs. (5) and (6).

$$p_{x,y} = \sqrt{\left(\boldsymbol{S}_{x,y} * \mathbf{G}_\mathrm{x}\right)^2 + \left(\boldsymbol{S}_{x,y} * \mathbf{G}_\mathrm{y}\right)^2}, \tag{5}$$

$$\theta_{x,y} = \tan^{-1}\left(\frac{\boldsymbol{S}_{x,y} * \mathbf{G}_\mathrm{y}}{\boldsymbol{S}_{x,y} * \mathbf{G}_\mathrm{x}}\right), \tag{6}$$

where $\boldsymbol{S}_{x,y}$ is the 3×3 neighborhood smoothed image.

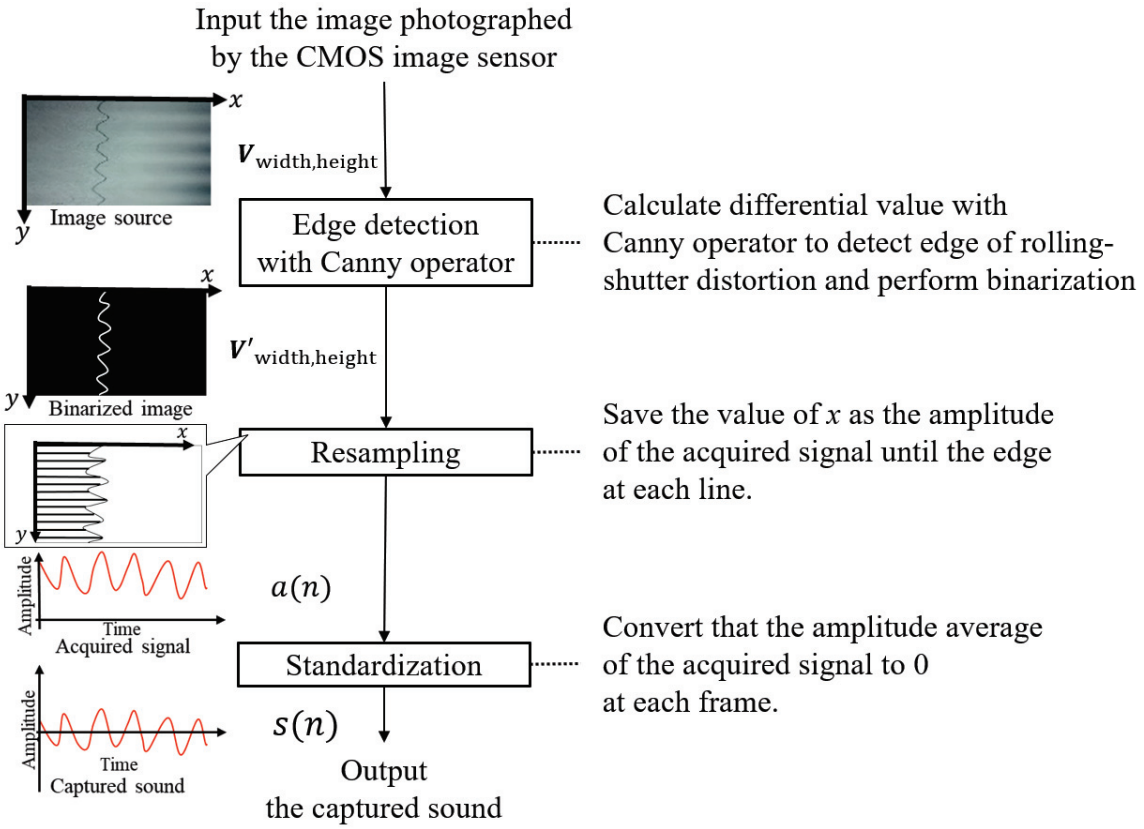


Figure 2 – Overview of the proposed method.

Step 3: Non-maximum suppression (NMS)

NMS is adopted to obtain the accurate positioning and refinement edge. If the target pixel's gradient is not maximal value among the three pixels adjacent in the gradient direction $\theta_{x,y}$, the gradient is set to 0. The gradient image adopted NMS $p'_{x,y}$ is given by Eq. (7).

$$\begin{aligned}
&\left(\text{Case A: } -\frac{1}{8}\pi \le \theta_{x,y} < \frac{1}{8}\pi \cup \theta_{x,y} > \frac{7}{8}\pi \cup \theta_{x,y} < -\frac{7}{8}\pi\right) \quad p'_{x,y} = \begin{cases} p_{x,y} & p_{x,y} > p_{x-1,y} \cap p_{x,y} > p_{x+1,y} \\ 0 & \text{otherwise} \end{cases}, \\
&\left(\text{Case B: } \frac{1}{8}\pi \le \theta_{x,y} < \frac{3}{8}\pi \cup -\frac{7}{8}\pi \le \theta_{x,y} < -\frac{5}{8}\pi\right) \quad p'_{x,y} = \begin{cases} p_{x,y} & p_{x,y} > p_{x+1,y-1} \cap p_{x,y} > p_{x-1,y+1} \\ 0 & \text{otherwise} \end{cases}, \\
&\left(\text{Case C: } \frac{3}{8}\pi \le \theta_{x,y} < \frac{5}{8}\pi \cup -\frac{5}{8}\pi \le \theta_{x,y} < -\frac{3}{8}\pi\right) \quad p'_{x,y} = \begin{cases} p_{x,y} & p_{x,y} > p_{x,y-1} \cap p_{x,y} > p_{x,y+1} \\ 0 & \text{otherwise} \end{cases}, \\
&\left(\text{Case D: } \frac{5}{8}\pi \le \theta_{x,y} \le \frac{7}{8}\pi \cup -\frac{3}{8}\pi \le \theta_{x,y} < -\frac{1}{8}\pi\right) \quad p'_{x,y} = \begin{cases} p_{x,y} & p_{x,y} > p_{x-1,y-1} \cap p_{x,y} > p_{x+1,y+1} \\ 0 & \text{otherwise} \end{cases},
\end{aligned} \tag{7}$$

Step 4: Hysteresis threshold

The hysteresis threshold is adopted to check and connect edges. The hysteresis threshold binarizes the image with two thresholds $T_{\min}$ and $T_{\max}$. If the gradient $p'_{x,y}$ is larger than $T_{\max}$, its pixel value $v'_{x,y}$ is set to 255 as the edge. If it is smaller than $T_{\min}$, its pixel value is set to 0 as the non-edge. The pixels that gradient $p'_{x,y}$ is between the two thresholds are set to 255 as long as there is at least one edge pixel around it; otherwise they are set to 0. From these steps, a binarized image $\boldsymbol{V}'$ is obtained by Canny operator. The proposed method adopts resampling and standardization to capture sound from the binarized image $\boldsymbol{V}'$.

Resampling

In the binarized image $\boldsymbol{V}'$, x coordinates indicate displacement of surface vibrations and y coordinates indicate time index n of the acquired signal. Thus, the proposed method stores x coordinates to edge of each row sequentially as the amplitude of the acquired signal. Then, resampling is performed according to sampling frequency of the acquired signal and that of the CMOS image sensor. The acquired signal $a(n)$ is given by Eq. (8).

$$a(n) = \arg\min_{x \in \boldsymbol{C}} x \quad \boldsymbol{C}: v'_{x,y} = 255, y = \frac{F_{\text{CMOS}}}{F_{\text{out}}} \times n \quad \left(0, \dots, n, \dots, \frac{F_{\text{out}}}{F_{\text{CMOS}}} \times \text{height}\right), \tag{8}$$

where n is time index of the acquired signal, F_{out} is the sampling frequency of the acquired signal, F_{CMOS} is that of CMOS image sensor, $v'_{x,y}$ is a pixel value of the binarized image $\boldsymbol{V}'$, and height is max value of y coordinate.

Standardization

The proposed method can capture acquired signal from the image by resampling. However, as shown in Fig. 2, all amplitudes of the acquired signal are positive values because these are calculated as the amount of deviation from the y-axis of the image. Therefore, it is necessary to convert the amplitude average to 0. The proposed method divides the acquired signal into short time frames and performs standardization so that the amplitude average of each frame is 0. By standardizing for each short time frame, it reduces low frequency noise included due to factors such as paper inclination. The lth frame of the captured sound $s(n,l)$ is given by Eq. (9).

$$s(n,l) = a(n,l) - \bar{a}(l), \tag{9}$$

where $\bar{a}(l)$ is average of the acquired signal at lth frame. We can obtain the sound that vibrates the object as the captured sound $s(n)$ by connecting frame $s(n,l)$.

4. Sound capture experiment

4.1 Experimental conditions

We carried out an experiment to evaluate the performance of the proposed method. The experimental arrangement and vibrating object are shown in Fig. 3. The experimental conditions are shown in Table 1. In this experiment, we used a half colored paper of A4-size as a vibrating object, and the boundary of the colored paper was located in 10 mm front of the loudspeaker. In addition, in order to obtain sufficient illuminance, a light projector was used as a light source because a CMOS image sensor in a dark place generates a lot of noise in the image. The experiment was conducted under the light source. As sound sources, pure tone sounds with 400 Hz, 1,000 Hz, and 1,600 Hz were respectively used and emitted from the loudspeaker. In this experiment, the height of each image source is 1,080 pixels and the sampling rate of the CMOS image sensor is 48,000 Hz. This means the proposed method can capture the sound of 25 ms length from one image source.

4.2 Experimental result

Image sources and binarized images are respectively shown in Figs. 4 and 5. In this section, for simply explanation, image sources and binarized images are displayed rotated 90 degrees and trimmed. Thus, the vertical direction indicates the displacement of the vibrating object and the horizontal direction indicates the time information by the CMOS image sensor. From Fig. 4, the amplitude of vibration affected by higher frequency sound is smaller than that of lower frequency sound. That is because an object with mass has a small displacement by a higher frequency sound. For this reason, from Fig. 5, edge detection accuracy of higher frequency sound is lower than that of lower frequency sound. It is considered that the captured sound distorts due to low edge detection accuracy.

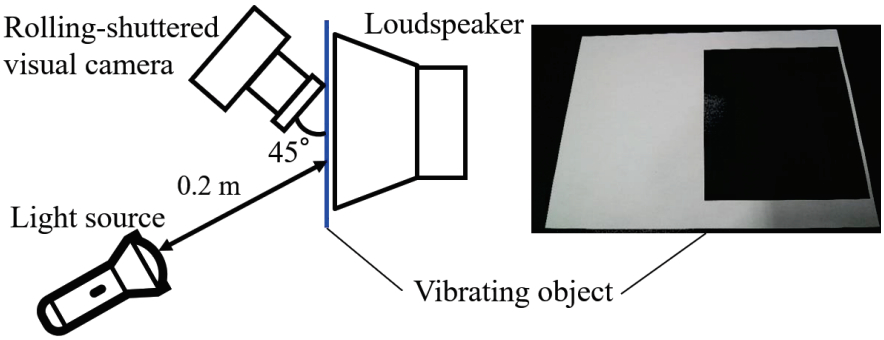
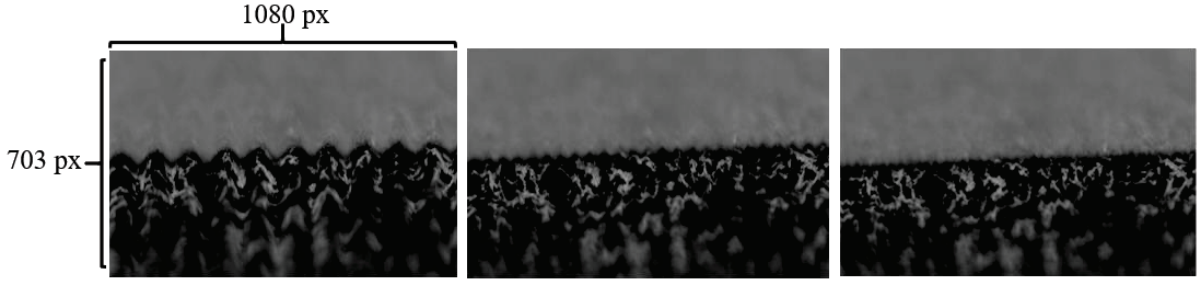


Figure 3 – Experimental arrangement and vibrating object.

Table 1 – Experimental conditions.

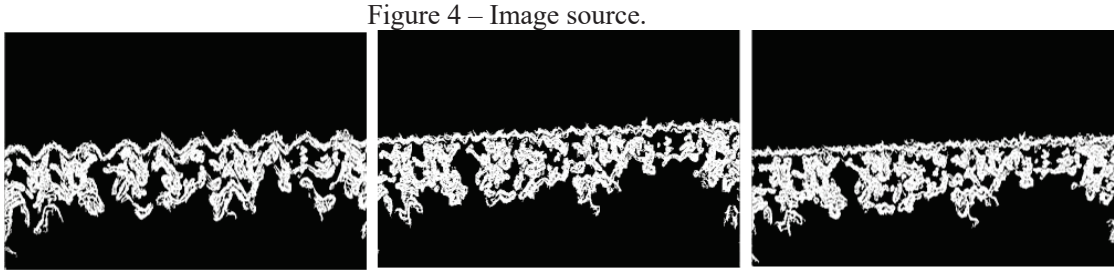
Environment	Soundproof room ($T_{60} = 100$ ms)
Background noise	$L_{\mathrm{A}} = 27.5$ dB
Temperature / Humidity	19.0 ℃ / 37.6 %
Sound source	Pure tone (400 Hz, 1,000 Hz, 1,600 Hz)
Sound pressure level	$L_a = 110$ dB at 0 m
Sampling frequency / bitrate	8,000 Hz / 16 bits
Rolling-shuttered camera	Canon, EOS 5D Mark II
Lens	Canon, MP-E 65 mm f/2.8 1-5x
Size of image source	1,920 × 1,080 px
Light source	LED light (10,000 lm)

Waveforms and power spectra of each captured sound are respectively shown in Figs. 6 and 7. From Fig. 6, as we expected, distortion of higher frequency sound is larger than that of lower frequency sound. From Fig. 7, we can confirm the peak at contained frequency of each sound. In other words, the proposed method can capture the sound from the image source. Furthermore, the SNR of the captured sound with higher frequency sound is worse than that of lower frequency sound due to distortion of its waveform. In particular, for the pure tone with 1,600 Hz, the power of the captured sound is lower than that of noise. For that reason, when capturing sound which contains frequencies above 1,600 Hz, it is necessary to use a high-zoom camera or a lens to photograph the sufficient displacement for detecting edges and reducing distortion. In addition, it is considered that performing weighted summation at multiple locations for edge detection is useful to emphasize the higher frequency sound. From these results, the proposed method can capture sound from the image photographed by the rolling-shuttered visual camera and the problems are clarified.



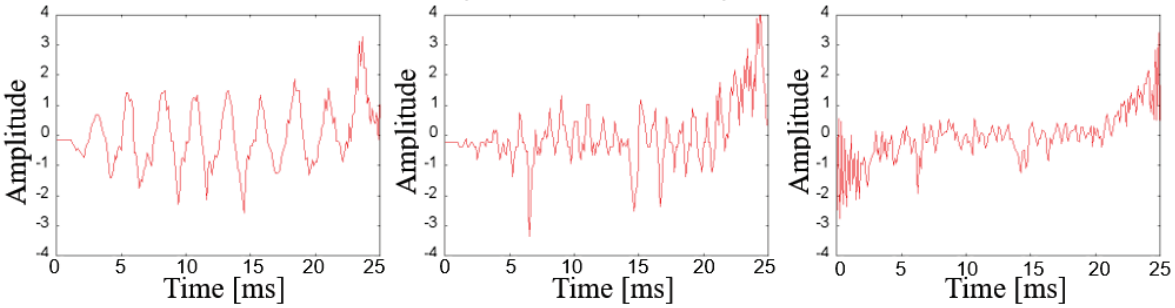
(a) 400 Hz (b) 1,000 Hz (c) 1,600 Hz

Figure 4 – Image source.



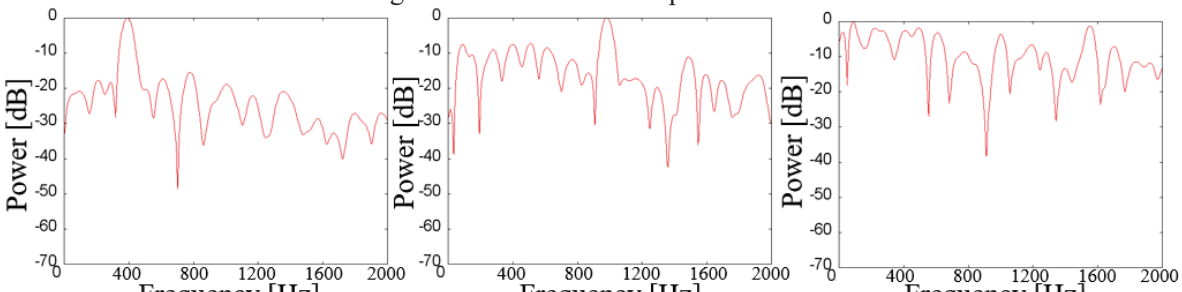
(a) 400 Hz (b) 1,000 Hz (c) 1,600 Hz

Figure 5 – Binarized image.



(a) 400 Hz (b) 1,000 Hz (c) 1,600 Hz

Figure 6 – Waveform of captured sound.



(a) 400 Hz (b) 1,000 Hz (c) 1,600 Hz

Figure 7 – Power spectra of captured sound.

5. CONCLUSIONS

In this paper, we proposed the new sound capture method from the image of the vibrating object affected by sound waves with a rolling-shuttered visual camera. The proposed method uses the edge detection to emphasize the edge of rolling-shutter distortion caused by sound waves, and measures displacement of vibration to capture the sound affecting the object. We conducted an experiment to confirm the performance of the proposed method. From the experimental results, we confirmed that the proposed method can capture pure tone sound with 1,000 Hz from the image photographed by the rolling-shuttered camera. In addition, we also confirmed that extracting higher frequency sound is more difficult than lower frequency sound due to small displacement of vibration. In the future, we intend to utilize acoustics signal processing to emphasize higher frequency sound and movie sources to capture various sounds.

ACKNOWLEDGEMENTS

This work was partly supported by JST-COI and JSPS KAKENHI Grant Numbers JP18K19829, JP19H04142, and R-GIRO (Ritsumeikan Global Innovation Research Organization) funded by Ritsumeikan University.

REFERENCES

1. J. Shan, Y. He, D. Liu and W Chen, "Laser Doppler Vibrometer for long-distance acoustical signals acquirement," *Conference on Lasers & Electro Optics & The Pacific Rim Conference on Lasers and Electro-Optics*, pp. 1-2, 2009.
2. M. Hua, L. Zhou, C. Liu and Z. Li, "The research of vibration detection using the visual microphone technology," *10th International Conference on Measuring Technology and Mechatronics Automation*, pp. 256-258, 2018.
3. J. Ahn and D. Kim, "Simple and effective speech enhancement for visual microphone," *4th IAPR Asian Conference on Pattern Recognition*, pp. 694-699, 2017.
4. S. Baker, E. Bennett, S. B. Kang and R. Szeliski, "Removing rolling shutter wobble," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2392-2399, 2010.
5. J. Chun, H. Jung and C. Kyung, "Suppressing rolling-shutter distortion of CMOS image sensors by motion vector detection," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1479-1487, 2008.
6. J. Canny, "A computation approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679-698, 1986.
7. A. P. Thombare and S. B. Bagal, "A distributed canny edge detector: Comparative approach," *International Conference on Information Processing*, pp. 312-316, 2015.
8. C. Sun, X. Fan and D. Zhao, "A fast intra cu size decision algorithm based on Canny operator and SVM classifier," *25th IEEE International Conference on Image Processing*, pp. 1787-1791, 2018.