Live Demonstration: Event-based Visual Microphone

Ryogo Niwa[1], Tatsuki Fushimi[1], Kenta Yamamoto[1], and Yoichi Ochiai[1,2]
[1] University of Tsukuba, [2] R&D Center for Digital Nature
{niwa.ryogo, tfushimi, kenta.yam,wizard}@digitalnature.slis.tsukuba.ac.jp

1. Experiment Setup

We used an event camera (SilkyEvCam HD) with a TAMRON AF180mm F3.5 Di lens to capture video of a fast-moving object in a general room lighting environment without any special lighting setup.

We also used a high-speed camera (Photron FASTCAM Nova R2) to capture video of a fast-moving object at 8000 fps. A NANLITE FS-300 light source was used for illumination. The lens used for the high-speed camera was a TAMRON lens, the same as the high-speed camera. As shown in the supplementary video, we used a SIGMA 150-600mm F6.3 lens with a SIGMA TELECONVERTER (TC-2001) to capture footage of an object at a distance of about 10 meters.

2. Method

2.1. Event-based Visual Microphone

2.1.1 Parameters

We provide the parameters used to restore the sound, which were $n = 13$, $\sigma = 3$, $\lambda = 32$, $\gamma = 1$, kernel $\phi = 0$ for the Gabor kernel and $\sigma = 8$ for the Gaussian kernel. And we set the parameters of the Gabor Filter according to the direction of the vibration. For example, $\theta = 0$ for vertical vibration and $\theta = \pi 2$ for horizontal vibration.

2.1.2 Signal Processing

We reconstructed the sound from about 5 pixels in descending order of the number of events. One of them was selected and shown as the experimental result of the paper. You can listen to it in the supplementary information wav file.

If we want to reconstruct the vibrations with greater accuracy, we can use Principal Component Analysis. First, we extract the vibration data from the N pixels with the highest event counts and create a matrix $\delta \in \mathbb{R}^{N \times T}$, where T is the number of samples in the recovered data. We perform a singular value decomposition on δ and compute $\eta = U_r * \delta$ to linearly project the vibration data into a few principal components, where U_r is the matrix consisting of the first r columns of U, i.e. $[u_1, ..., u_r] \in \mathbb{R}^{N \times r}$. The ith row of $\eta = [\eta_1, ..., \eta_r] \in \mathbb{R}^{r \times T}$ corresponds to the i th principal component of δ, thus providing a more precise signal representation.

2.2. High-Speed Camera

The method of extracting sound from video footage captured by a high-speed camera is known as 'visual microphone', and is described in detail in the paper by Abe et al [1]. The code used in this study can be found at this GitHub link: https://github.com/dsforza96/visual-mic.

3. Result

The restored audio we obtained is available in the supplementary materials. These materials include both the audio recorded by the microphone and the audio restored using our method. The audio files used in Figures 2(b), (c), and (d) of the paper correspond to "6-1.wav" for Figure 2(b), "6-2.wav" for Figure 2(d), and "6-3.wav" for Figure 2(c), respectively. In addition, Figure 2(e) corresponds to "3-1.wav" in the paper, Figure 2(f) corresponds to "3-3.wav", and Figure 2(g) corresponds to "3-2.wav".

Sounds with PCA in the filename used the method described in Sec. 2.1.2. This sound uses the sound restored from 3000 pixels and then subjected to the signal processing described above.

References

[1] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics*, 33(4):1–10, July 2014. 1