Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Audio extraction from silent high-speed video using an optical technique

Zhaoyang Wang
Hieu Nguyen
Jason Quisberth

SPIE.

Audio extraction from silent high-speed video using an optical technique

Zhaoyang Wang,* Hieu Nguyen, and Jason Quisberth
The Catholic University of America, Department of Mechanical Engineering, Washington, DC 20064, United States

Abstract. It is demonstrated that audio information can be extracted from silent high-speed video with a simple and fast optical technique. The basic principle is that the sound waves can stimulate objects encountered in the traveling path to vibrate. The vibrations, although usually with small amplitudes, can be detected by using an image matching process. The proposed technique applies a subset-based image correlation approach to detect the motions of points on the surface of an object. It employs the Gauss-Newton algorithm and a few other measures to achieve very fast and highly accurate image matching. Because the detected vibrations are directly related to the sound waves, a simple model is introduced to reconstruct the original audio information of the sound waves. The proposed technique is robust and easy to implement, and its effectiveness has been verified by experiments. © *2014 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.OE.53.11.110502]

Keywords: audio extraction; high-speed video; sound vibration; image matching; digital image correlation.

Paper 141476L received Sep. 19, 2014; accepted for publication Oct. 29, 2014; published online Nov. 21, 2014.

In recent years, the steady development and growth of camera technology have made high-speed imaging more affordable and prevalent. Meanwhile, image matching techniques have gained significant improvement and tremendous popularity in optics and a variety of other fields. A combination of high-speed imaging techniques and image matching techniques has recently resulted in numerous unprecedented applications, particularly in vibration and dynamic measurements.[1–5]

Sound waves are mechanical waves that cause the medium (e.g., air) to vibrate when traveling. The vibration through air can further lead to the vibration of objects that are located in its traveling path, especially if the objects are lightweight, thin, and flexible (e.g., a piece of paper). The surface vibration of an object can traditionally be measured by using a vibration sensor such as a laser Doppler vibrometer. Very recently, a report on using video to recover sound has drawn remarkable public and media attention,[6] where the authors used a sophisticated technique that involves decomposing video into complex-valued spatial subbands to compute motion signals. With the aforementioned image matching approach for vibration measurement, this letter aims to present a simpler and much faster technique that can extract audio information from a silent high-speed video. Because light can travel through air considerably farther than sound and it can pass through glass, the technique may find useful applications as technologies continue to evolve.

Image matching is also called digital image correlation (DIC) in the fields of optics and mechanics and image registration in computer vision and other fields.[7–9] The DIC technique normally uses artificial speckle patterns and puts emphasis on accuracy, whereas the imaging registration technique often demands more speed and the convenience of using natural patterns.[10] In this work, both a high accuracy and the usage of natural patterns are essential.

When an object vibrates following sound waves, the magnitudes of the surface displacements are generally small. Consequently, the corresponding pixel displacements or disparities between two different images are usually less than one pixel. This requires the image matching process to be capable of mapping the points in one image (the reference) to their corresponding positions in another image (the target) with subpixel accuracies (e.g., 0.01 pixels). To provide necessary features and sufficient information for the matching, the object's surface should have varying texture intensities.

To start the processing, a region of interest (ROI) in the reference image where matching will be conducted is specified first. Next, the ROI is divided into evenly spaced virtual grids (grid distance can be set to 10 pixels or larger), so that the displacements will be calculated at each point of the virtual grid instead of every pixel location to substantially reduce the processing time without sacrificing accuracy. Then, the process employs a correlation criterion to detect the matching for a set of pixels (named a subset) centered at each grid pixel to be interrogated, and this can be written as[11]

$$C = \sum_{i=1}^{N} [af(x_i, y_i) + b - g(x_i', y_i')]^2, \tag{1}$$

where N is the number of pixels in the subset, a is a scale factor, b is an offset of intensity, and $f(x_i, y_i)$ and $g(x_i', y_i')$ indicate the intensity values at the i'th pixel in the reference subset and the potential matching pixel in the target subset, respectively. The task of the correlation analysis is to minimize the coefficient C in Eq. (1) to detect the best matching. For a grid pixel (x_0, y_0) to be analyzed in the reference image, a square pattern of $N = (2M+1) \times (2M+1)$ pixels with its center located at (x_0, y_0) is often selected as the reference subset. The corresponding subset in the target image, i.e., the target subset, is usually of irregular shape. Denoting the translation amount between the centers of the two potential matching subset patterns as (ξ, η), a shape mapping function for the entire reference and target subsets can be expressed as

$$\begin{aligned} x_i' &= x_i + \xi + \xi_x \Delta x_i + \xi_y \Delta y_i, \\ y_i' &= y_i + \eta + \eta_x \Delta x_i + \eta_y \Delta y_i, \end{aligned} \tag{2}$$

where $i = 1, 2, \ldots, N$; $\Delta x_i = x_i - x_0$, $\Delta y_i = y_i - y_0$, and coefficients ξ_x, ξ_y, η_x, and η_y represent the displacement gradients. To determine all six unknowns of the shape function as well as the scale and offset parameters involved in Eq. (1), an

*Address all correspondence to: Zhaoyang Wang, E-mail: wangz@cua.edu

iterative approach can be used to carry out the correlation optimization. To better show the iterative algorithm, Eq. (1) is rewritten in another form as

$$C(\boldsymbol{p}) = \sum_{i=1}^{N} [\zeta_i(\boldsymbol{p})]^2, \tag{3}$$

where vector $\boldsymbol{p} = \{\xi, \eta, \xi_x, \xi_y, \eta_x, \eta_y, a, b\}^\intercal$ contains all the unknown parameters, and $\zeta_i(\boldsymbol{p}) = af(x_i, y_i) + b - g(x_i', y_i')$. The best estimate of the mapping parameters is established by minimizing $C(\boldsymbol{p})$. This can be iteratively carried out by applying the Gauss-Newton algorithm[12] to Eq. (3), which yields the governing equation as

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_n - \left[\sum_{i=1}^{N} (\boldsymbol{J}_i \cdot \boldsymbol{J}_i^\intercal)\right]^{-1} \cdot \sum_{i=1}^{N} [\zeta_i(\boldsymbol{p}) \cdot \boldsymbol{J}_i]. \tag{4}$$

In the equation, $n = 0, 1, 2, \ldots$; $\boldsymbol{J}_i = [\partial \zeta_i(\boldsymbol{p})/\partial \boldsymbol{p}] = -\{g_{x_i'}, g_{y_i'}, g_{x_i'}\Delta x_i, g_{x_i'}\Delta y_i, g_{y_i'}\Delta x_i, g_{y_i'}\Delta y_i, -f_i, -1\}^\intercal$ is the Jacobian vector, where f_i denotes $f(x_i, y_i)$; $g_{x_i'} = [\partial g(x_i', y_i')/\partial x_i']$ and $g_{y_i'} = [\partial g(x_i', y_i')/\partial y_i']$ are the intensity gradients of the target subset at location (x_i', y_i') in the x- and y-directions, respectively. The initial guess for starting the iteration process is $\boldsymbol{p}_0 = \{0, 0, 0, 0, 0, 0, 1, 0\}^\intercal$. This simple yet effective initial guess is due to the small displacements involved in the vibration measurement.

The iteration of Eq. (4) requires obtaining the intensities at noninteger-pixel or subpixel locations in the target images. Therefore, a subpixel interpolation algorithm must be used such as the bicubic interpolation function or the fast recursive B-spline interpolation functions.[13] In the iteration, the convergence tolerance can be set to 1×10^{-4} for each element of $\boldsymbol{p}$, and it usually takes only a few cycles to converge. To deal with possible iteration divergence, a threshold can be specified to the maximum allowable number of iteration cycles, such as 30. If the iterations do not converge, then the matching at the corresponding grid pixel is regarded as a failure and will be excluded from subsequent processing.

Each subset matching process gives a result of $\boldsymbol{p}$ for the subset pair, but only ξ and η are eventually used because they represent the horizontal and vertical displacements, respectively, at the subset center (i.e., a grid pixel point).

It is reasonable to assume that the vibration amplitude of the object is proportional to the amplitude of the original audio waves. In addition, data averaging is adopted to reduce the random error and to yield a single value for each captured image. Consequently, the amplitude of the audio wave at a sample point can be reconstructed as

$$A = \frac{1}{S} \sum_{j=1}^{S} \sqrt{\xi_j^2 + \eta_j^2}, \tag{5}$$

where S denotes the number of valid grid pixels, where image matching has been successfully completed. In general, the amplitude A should be filtered and normalized to a range of $[-1, 1]$ for audio reconstruction.

The proposed approach selects one image as the target image and all others as the reference images. Such handling allows the generation of coefficients for the aforementioned interpolation function to be conducted only once, and this helps to substantially reduce the processing time. In practice, the target image can be selected from those captured at the time when there is no or relatively weak vibration. Another useful scheme to enhance the processing speed without perceptible loss of accuracy is to use every other pixel in the x- and y-directions in the subset during intermediate iteration.

The proposed technique can be summarized in four key steps. First, capture the images of a thin and flexible object using a high-speed camera. Second, select target and reference images and specify processing parameters (ROI, subset size, and so on) to analyze the images using Eq. (4). Third, determine the vibration displacements and calculate the mean values using Eq. (5). Finally, reconstruct the audio using the mean displacement of each image.

To demonstrate the validity and effectiveness of the proposed technique, many experiments have been conducted in this work. An example is presented below.

In the experiment, a piece of newspaper is attached to an optical post as the object to catch the sound waves, which are generated by a speaker placed near the paper. A Phantom V310 high-speed camera is employed to capture a small region on the newspaper with a resolution of 256×256 pixels at 1000 frames per second (fps) while an audio clip is played. Although a higher frame rate is desired, 1000 fps is adopted in the measurement to consider the trade-off between measurement details and computation time. Furthermore, no apparent distinction can be found in the reconstructed audio quality using frame rates higher than 1000 fps. Figure 1 illustrates some key steps and the results of the processing. For the purpose of clear demonstration, Figs. 1(b)–1(d) show the full-field pixel disparities in the ROI, although the results only need to be determined at the grid pixels.

In the image matching analysis, the subset size is set to 21×21 pixels and the grid distance is set to 15 pixels.

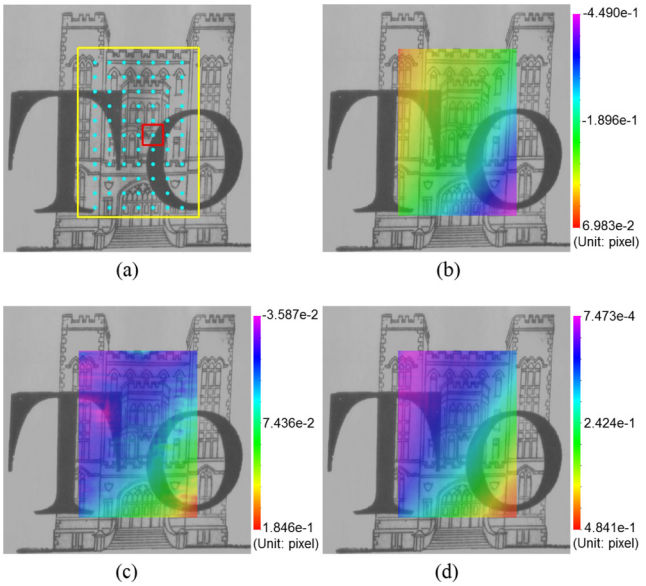

Fig. 1 Key steps and results in an experiment. The background image in (a) is an arbitrarily selected reference image, and the background image in (b)–(d) is the target image. (a) ROI and analysis grids as well as the reference subset for a randomly chosen grid pixel, (b) map of detected pixel disparity in x-direction, (c) map of detected pixel disparity in y-direction, and (d) magnitude map of resultant pixel disparity.

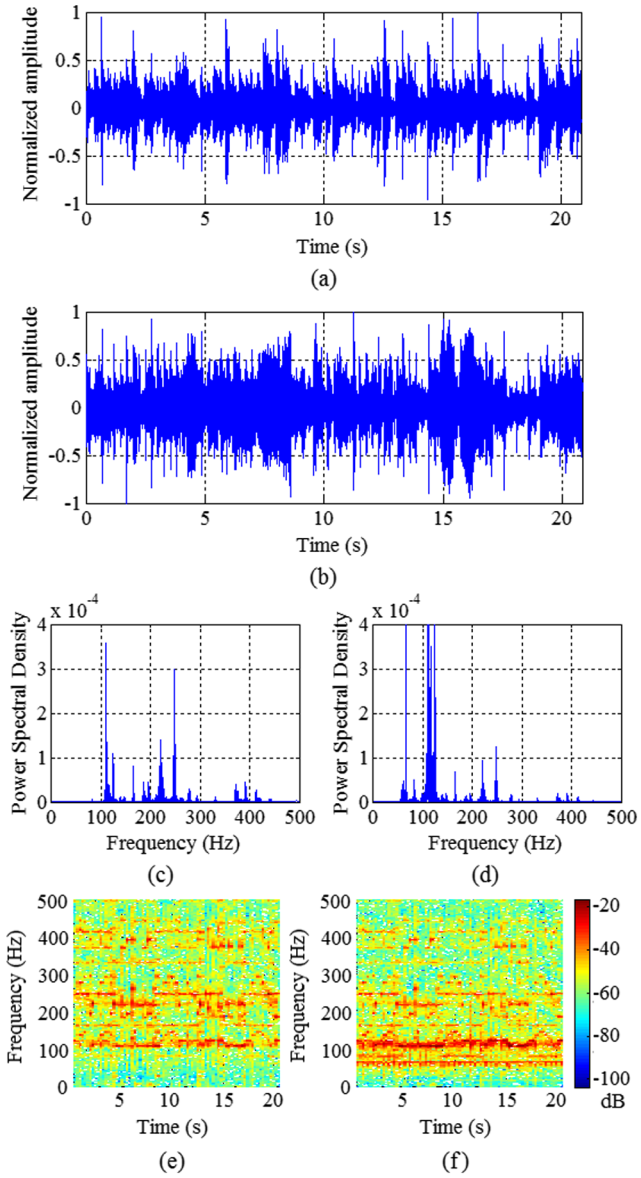

Fig. 2 Experimental result and comparisons. (a) Original audio waveform, (b) reconstructed audio waveform, (c) power spectral density of original audio, (d) power spectral density of reconstructed audio, (e) original audio spectrogram, and (f) reconstructed audio spectrogram.

The processing time for the 21-s video or 21,000 images is within 3 min with a regular laptop computer (an Intel i7 processor and 4 GB memory). For the specific image shown in Fig. 1(a), the processing gives an amplitude value of 0.1606 pixels. Figure 2 shows the comparisons between the original and reconstructed sounds in the forms of audio waveforms, power density distributions, and spectrograms. It can be seen that the reconstructed audio matches the original quite well. This has been confirmed by directly listening to the two audio clips. Figure 2 also reveals the existence of obvious errors. The errors are due to a number of reasons including the stronger response of the object to low-frequency audio waves than high-frequency ones, the oversimplified relationship between the inplane motion and the audio waveform, the limited sensitivity to tiny vibrations, the inherent noise involved in high-speed imaging, and so on. It should be noted that a desired and quantitative way to compare the two audio signals is to perform spectral coherence analysis, which is technically similar to the image matching process employed in this work. However, because the proposed technique is incapable of capturing all the details of the original audio signals and because the noise level of the result is high, a spectral coherence analysis is not suitable for this work. As the proposed technique is being improved, a spectral-coherence-based comparison would be feasible in the future. It is also noteworthy that the processing time can be further reduced to below 2 or 1 min by using a larger grid distance and/or a smaller ROI, which, however, can attenuate the processing reliability. The spectral coherence analysis will also be helpful for the investigation of optimized grid distance and ROI size.

Technically, it is favorable to use two video streams captured by two stereo-vision-positioned cameras[14] for audio extraction because vibrations can be measured in three dimensions. In practice, this approach is challenging because of the difficulty in synchronizing high-speed cameras. Nevertheless, this can be a future work along with the tasks on enhancing the sensitivity of measurement and investigating a rigorous relationship between the detected vibrations and the original audio.

Acknowledgments

This work was supported by the U.S. Army Research Office (ARO) under grant W911NF-10-1-0502.

References

1. M. Pankow, B. Justusson, and A. Waas, "Three-dimensional digital image correlation technique using single high-speed camera for measuring large out-of-plane displacements at high framing rates," *Appl. Opt.* **49**, 3418–3427 (2010).
2. M. Helfrick et al., "3D digital image correlation methods for full-field vibration measurement," *Mech. Syst. Signal Process.* **25**, 917–927 (2011).
3. J. Sirohi and M. Lawson, "Measurement of helicopter rotor blade deformation using digital image correlation," *Opt. Eng.* **51**, 043603 (2012).
4. F. Trebuna and M. Hagara, "Experimental modal analysis performed by high-speed digital image correlation system," *Measurement* **50**, 78–85 (2014).
5. R. Wu et al., "Determination of three-dimensional movement for rotary blades using digital image correlation," *Opt. Laser. Eng.* **65**, 38–45 (2015).
6. A. Davis et al., "The visual microphone: passive recovery of sound from video," *ACM Trans. Graphics* **33**, 79 (2014).
7. L. Yu and B. Pan, "In-plane displacement and strain measurements using a camera phone and digital image correlation," *Opt. Eng.* **53**, 054107 (2014).
8. J. Salvi et al., "A review of recent range image registration methods with accuracy evaluation," *Image Vision Comput.* **25**, 578–596 (2007).
9. F. Oliveira and J. Tavares, "Medical image registration: a review," *Comput. Methods Biomech.* **17**, 73–93 (2014).
10. Z. Wang et al., "Digital image correlation in experimental mechanics and image registration in computer vision: similarities, differences and complements," *Opt. Laser. Eng.* **65**, 18–27 (2015).
11. B. Pan, H. Xie, and Z. Wang, "Equivalence of digital image correlation criteria for pattern matching," *Appl. Opt.* **49**, 5501–5509 (2010).
12. J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer, New York (2006).
13. L. Luu et al., "Accuracy enhancement of digital image correlation with B-spline interpolation," *Opt. Lett.* **36**, 3070–3072 (2011).
14. H. Kieu et al., "Accurate 3D shape measurement of multiple separate objects with stereo vision," *Meas. Sci. Technol.* **25**, 035401 (2014).