Dual-Shutter Optical Vibration Sensing

Mark Sheinin, Dorian Chan, Matthew O’Toole, and Srinivasa G. Narasimhan
Carnegie Mellon University, Pittsburgh, PA 15213, USA
marksheinin@gmail.com, dychan@andrew.cmu.edu, mpotoole@cmu.edu, srinivas@andrew.cmu.edu

Abstract

Visual vibrometry is a highly useful tool for remote capture of audio, as well as the physical properties of materials, human heart rate, and more. While visually-observable vibrations can be captured directly with a high-speed camera, minute imperceptible object vibrations can be optically amplified by imaging the displacement of a speckle pattern, created by shining a laser beam on the vibrating surface. In this paper, we propose a novel method for sensing vibrations at high speeds (up to 63kHz), for multiple scene sources at once, using sensors rated for only 130Hz operation. Our method relies on simultaneously capturing the scene with two cameras equipped with rolling and global shutter sensors, respectively. The rolling shutter camera captures distorted speckle images that encode the high-speed object vibrations. The global shutter camera captures undistorted reference images of the speckle pattern, helping to decode the source vibrations. We demonstrate our method by capturing vibration caused by audio sources (e.g. speakers, human voice, and musical instruments) and analyzing the vibration modes of a tuning fork.

1. Introduction

Vibrations are all around us, caused by sources ranging from heartbeats to engines, to music, speech, and ultrasonics. These vibrations exhibit various amplitudes (microns to meters) and frequencies (a few Hz to MHz). As such, measuring vibrations is an essential tool in many engineering and scientific fields. However, optically sensing vibrations, especially the low-amplitude high-frequency kind, is challenging. To make matters worse, indirect damped vibrations caused by remote sources (e.g. a speaker vibrating an object [13]) can be even more subtle. Additionally, these challenges are harder to overcome when the vibrating surface is far from the imaging system or is itself moving (e.g. the natural movements of a musician playing guitar).

That said, much progress has been made recently on visual vibrometry. Passive capture and estimation of small motions [15, 18, 21, 35–37, 39, 45] of a vibrating surface have been used to extract (a) heart rate and sound from video [13, 39] and (b) the physical properties of materials [6,9,12,16] and structures (e.g. buildings and bridges) [10]. However, low-amplitude high-frequency vibrations require a high-speed video camera with a zoom-lens and bright lighting to compensate for the short camera exposures.

In contrast to passive approaches, active speckle-based approaches illuminate a vibrating surface with coherent light (e.g. laser) and image the resulting speckle [5, 40–42]. The speckle is imaged by focusing in between the surface and the sensor. A small tilt of the vibrating surface results in a shift of the speckle, a phenomenon called the memory effect [3, 42]. This approach optically magnifies small-amplitude vibrations and has been used to demonstrate long-distance audio capture [5, 40, 42]. But, as in

1The memory effect was used for motion tracking [22,47], ego-motion [17], object tracking [30,31], environment sensing [43,44] and more.
the passive case, high-frequency vibrations would require expensive 2D high-speed cameras, whose available bandwidth limits either the maximum sampling frequency or the captured video’s spatial resolution. In response, some works \cite{5,20,40,41} use fast 1D sensors to capture high frequencies, but they can only reconstruct vibrations along one dimension.

We present a novel imaging system that exploits the speed of a 1D sensor but still estimates 2D speckle-based vibrations at high frequencies in a bandwidth-efficient manner. We are inspired by many works \cite{2,4,11,23,26,29,38} that use a rolling-shutter sensor as a 1D sensor to achieve imaging at high speeds. With this observation, we make two important changes to conventional speckle-based vibration imaging. First, we add a cylindrical lens to spread the speckle image to cover the entire vertical field of view of the rolling-shutter sensor. This allows us to extend speckle-based methods to sample multiple vibration locations simultaneously, using a low-speed camera, for the first time.

However, the captured speckle is distorted by the rolling shutter with unknown shifts at each image row. This distortion makes it hard to recover the 2D vibrations other than in very specialized situations (specific object motion, texture, and camera viewpoint resulting in horizontal vibration) \cite{12}. Thus, we propose using a second co-located low-speed 2D global-shutter camera that serves two purposes: tracking the appearance of the undistorted speckle pattern at low frequencies and providing a reference for recovering the high-frequency 2D shifts of the rolling-shutter sensor’s rows. We present an algorithm to recover the unknown high-speed shifts. This algorithm provides both the speckle pattern’s macro-motion or drift, along with the high-frequency vibrations.

Our dual-shutter system can recover vibrations with a range of amplitudes and frequencies (up to 63 kHz) using two ‘slow’ cameras (60 and 134 Hz).\footnote{Zhong et al. \cite{46} use a similar rolling-global shutter pair to generate ground truth frames for rectifying and deblurring rolling-shutter video.} We evaluate the system by (a) estimating the different known vibration modes of a tuning fork and (b) recovering high-quality audio by observing the membrane of a speaker. Since the system can simultaneously capture multiple vibrating surfaces, we demonstrate the separation of audio signals from multiple sources (e.g., musical instruments). The system can also capture subtle, indirect vibrations of a surface reacting to a nearby sound source in better quality than passive approaches. To measure speckle reliably with low-power lasers, we attach retroreflective markers on the vibrating surfaces. Our approach works without markers when the vibrating surfaces are near and not dark. We believe our system makes visual vibrometry of complex high-frequency vibrations more efficient and practical for many applications.

We make two changes to the standard speckle imaging described above as is illustrated in Fig. 2(b). First, the speckle focus-plane image is split using relay optics and imaged by two separate sensors having a rolling and a global shutter, respectively. Second, a cylindrical lens is placed in front of the primary objective lens. This lens spreads the speckle pattern along the vertical direction, yielding a ‘speckle column’ that reaches all the rolling-shutter sensor rows, while occupying only a fraction of the sensor’s image columns (e.g., 150 pixels). Therefore, unlike defocusing a ‘conventional’ spherical lens or using a bare sensor, our optical design can sample multiple surface points at once. Each point yields a separate speckle column that is sampled using all rolling-shutter rows (see Fig. 1).

Suppose that the rolling- and global-shutter sensors have
identical resolutions and exposures in the two sensor planes (blue rectangles in Fig. 2(b)). Then, since both sensors share the optical path, an identical image is formed on both sensors. Let \( I(x, t) \) be the image intensity in both sensors, where \( x \equiv (x, y) \) is the pixel coordinates and \( t \) is the image trigger time. For brevity, our equations below use both vector \( x \), and the explicit row coordinates \( y \) of \( x \). Note that \( I(x, t) \) is a continuous function of time, yielding the image, in [grayscale] units, that would form at trigger time \( t \).

Let \( I_{GS}(x, t) \), and \( I_{RS}(x, t) \) denote the global- and rolling-shutter video frames captured at time \( t \), respectively. In the global-shutter camera, all sensor pixels collect scene light simultaneously during the exposure duration, hence:

\[
I_{GS}(x, t) = I(x, t). \tag{1}
\]

Hereafter, we refer to the global-shutter frames as the reference frames. In a rolling-shutter sensor, the individual image rows are exposed one after another with a constant delay \( D \). Thus, the rolling-shutter frame at time \( t \) is:

\[
I_{RS}(x, t) = I(x, t + yD). \tag{2}
\]

Eqs. (1) and (2) describe the spatio-temporal relationship between the rolling-shutter and global-shutter videos:

\[
I_{RS}(x, t) = I_{GS}(x, t + yD), \quad y \in \{0, 1, ..., H - 1\}, \tag{3}
\]

where \( H \) is the number of rows in the rolling-shutter frame.

Now suppose that both cameras simultaneously start video capture at their individual frame rates. Let \( t_{k}^{GS} \) denote the time stamps of \( K \) global-shutter reference frames, where \( k = 0, 1, ..., K - 1 \) is the frame index. Similarly, let \( t_{n}^{RS} \) denote the time stamps of \( N \) rolling-shutter frames, where \( n = 0, 1, ..., N - 1 \) is the frame index (see Fig. 3).

As discussed above, for small tilts and shifts of the illuminated surface, the imaged speckle pattern remains approximately constant, up to a 2D image-domain shift

\[
u(t) \equiv (u_{dx}(t), u_{dy}(t)), \tag{4}
\]

where \( u_{dx}(t) \) and \( u_{dy}(t) \) are the x- and y-axis speckle pattern shifts in pixels, respectively. Without loss of generality, we set \( u(t_{0}^{GS}) = (0, 0) \). Thus, any two reference frames with indices \( k_{1} \) and \( k_{2} \) are related by image translation:

\[
I_{GS}(x, t_{k_{i}}^{GS}) = I_{GS}(x + u(t_{k_{1}}^{GS}) - u(t_{k_{2}}^{GS}), t_{k_{2}}^{GS}). \tag{5}
\]

Observe that the absolute shift \( u(t_{k}^{GS}) \) of any individual reference frame can be recovered by integrating all the relative image translations \( u(t_{k}^{GS}) - u(t_{k-1}^{GS}) \):

\[
u(t_{k}^{GS}) = \sum_{i=1}^{k} (u(t_{i}^{GS}) - u(t_{i-1}^{GS})) \quad \forall k > 0 \tag{6}
\]

\(^{3}\)For scenes having multiple points, Eq. (5) holds for each point’s speckle column individually.

Combining Eqs. (3)-(5) we get:

\[
I_{RS}(x, t_{n}^{RS}) = I_{GS}(x, t_{n}^{RS} + yD) = I_{GS}(x + u(t_{n}^{RS} + yD) - u(t_{k}^{GS}), t_{k}^{GS}). \tag{7}
\]

Let

\[
\delta u_{nk}(y) = u(t_{n}^{RS} + yD) - u(t_{k}^{GS}). \tag{8}
\]

denote the relative shift of every rolling-shutter row \( y \) in \( I_{RS}(x, t_{n}^{RS}) \) with respect to the same row in \( I_{GS}(x, t_{k}^{GS}) \). In Eq. (8) the term \( u(t_{k}^{GS}) \) is constant since all global-shutter frame rows are shifted together at time \( t_{k}^{GS} \). Rearranging Eq. (8) yields a formula for the speckle image shifts starting at time \( t_{n}^{RS} \) and ending at time \( t_{n}^{RS} + HD \):

\[
u(t_{n}^{RS} + yD) = \delta u_{nk}(y) + u(t_{k}^{GS}). \tag{9}
\]

Eq. (9) yields an important observation: given any pair of rolling- and global-shutter frames, we can compute \( H \) samples of the global speckle shifts with a fine temporal resolution of \( D \). All we require to recover the samples using Eq. (9) are two pieces of information: the shift \( u(t_{k}^{GS}) \), and \( \delta u_{nk}(y) \). The shift \( u(t_{k}^{GS}) \) can be computed using Eq. (6). In the next section, we discuss how to compute \( \delta u_{nk}(y) \).

3. Recovering the 2D speckle translation

In this section, we discuss how to recover speckle shifts \( \nu(t) \) using Eq. (9). Eq. (9) suggests that a single rolling-shutter frame, having index \( n \), can yield \( H \) temporal measurements of \( \nu(t) \) at a sampling rate of \( 1/D \). Therefore, given no delay between two consecutive rolling-shutter frames, capturing \( N \) consecutive frames yields a recording of \( NHD \) seconds duration.

Recovering \( \nu(t_{n}^{RS} + yD) \) requires selecting an appropriate reference frame \( k \). Note that in principle, any reference frame \( k \) should suffice. However, object macro motion may yield little to no spatial overlap between the speckle patterns of \( I_{RS}(x, t_{n}^{RS}) \) and \( I_{GS}(x, t_{k}^{GS}) \), causing the estimation of \( \delta u_{nk}(y) \) to fail. Therefore, one must select a reference frame whose timestamp \( t_{k}^{GS} \) is close to \( t_{n}^{RS} \).
First, \( I_{RS} \) and \( I_{GS} \) are cropped to the speckle column belonging to the object point we wish to recover (see Fig. 4(a) and (b)). Let \( I_{RS}(\mathbf{x}, t_n^{rs}) \) and \( I_{GS}(\mathbf{x}, t_n^{gs}) \) denote the resulting cropped videos. For brevity, let \( \hat{u}_{n,k}(y) \) denote the recovered shifts resulting from using reference frame \( k \):

\[
\hat{u}_{n,k}(y) \equiv \delta \hat{u}_{n,k}(y) + \hat{u}(t_n^{rs}),
\]

We use phase correlation to compute the shifts between every pair of consecutive reference frames \( \hat{I}_{GS}(\mathbf{x}, t_n^{gs}) \), and use Eq. (6) to yield \( \hat{u}(t_n^{rs}) \) \( \forall k \) [14].

Borrowing notation from [28], let \( \mathcal{V} = \{v_m\}_{m=0}^{M-1} \) denote a discrete set of \( M \) possible 2D row shifts, having some sub-pixel resolution and maximum span. \(^4\) Define the set of all row shifts for frame \( n \) as \( \mathcal{U} = \{\delta \hat{u}_{n,k}(y)\}_{y \in \mathcal{Y}} \), where \( \hat{u}_{n,k}(y) \in \mathcal{V} \). Then we recover \( \hat{U} \) by minimizing the loss:

\[
E(\hat{U}) = \sum_{y} [1 - S_y(\delta \hat{u}_{n,k}(y))] + \lambda \sum_{y,y'} V_{y,y'}(\delta \hat{u}_{n,k}(y), \delta \hat{u}_{n,k}(y')),
\]

where, the data term \( S_y(\delta \hat{u}_{n,k}(y)) \leq 1 \) quantifies the similarity of row \( y \) in \( I_{RS} \) to all \( M \) possible shifts of row \( y \) in \( \hat{I}_{GS} \). The term \( V_{y,y'}(\delta \hat{u}_{n,k}(y), \delta \hat{u}_{n,k}(y')) \) enforces smoothness by providing a penalty when neighboring rows \( y, y' \) have differing shifts [7]. We set \( V_{y,y'} = ||\delta \hat{u}_{n,k}(y) - \delta \hat{u}_{n,k}(y')||_2^2 \). We compute \( S_y(v_m) \) using the zero-normalized cross-correlation operator ZNCC(\( y \)) \[8\]:

\[
S_y(v_m) = \text{ZNCC} \left( \hat{I}_{RS}(\mathbf{x}, t_n^{rs}), \hat{I}_{GS}(\mathbf{x} + v_m, t_n^{gs}) \right).
\]

Finally \( \hat{U} \) is recovered using

\[
\hat{U} = \arg \min_{\hat{U}} \left( E(\hat{U}) \right). \tag{13}
\]

But solving Eqs. (11)-(13) directly for large \( M \) is computationally expensive, as it requires computing correlations with a large ‘dictionary’ of possible shifts. Thus, we additionally implement and use an efficient coarse-to-fine approach for solving Eq. (13) which computes correlations in the Fourier domain (see supplementary for details).

### 3.1. Merging multiple reference frames

In the above description, shift recovery for frame \( n \) relied on a single reference frame. However, as illustrated in Fig. 5, due to large amplitude vibrations or large object motions, a single reference may not be enough to recover the relative translations for all \( H \) rows, yielding partial recovery of \( u(t_n^{rs} + yD) \). Therefore, we use \( P \geq 1 \) reference frames to estimate \( \hat{u}(t_n^{rs} + yD) \), as described below.

Let \( \mathcal{R}_n = \{k_0, k_1, \ldots, k_{P-1}\} \) denote the set of indices of reference frames chosen to recover frame \( n \). For scenes having large low-frequency motions (e.g. hand-held instruments), \( \mathcal{R}_n \) consists of the \( P \) temporally closest frames to \( t_n^{rs} \). For mostly static scenes, we construct \( \mathcal{R}_n \) using frames close to \( t_n^{rs} \), that ‘cover’ the largest 2D speckle domain; see supplementary for details. First, shifts \( \hat{u}_{n,k}(y) \) are computed for every reference frame \( k \in \mathcal{R}_n \). Then, the shifts from all reference frames are merged using a weighted average:

\[
\hat{u}(t_n^{rs} + yD) = \sum_{k \in \mathcal{R}_n} W_{nk}(y) \hat{u}_{n,k}(y). \tag{14}
\]

Each reference frame’s per-row weights \( W_{nk}(y) \) are computed using the similarity measures of the recovered shifts:

\[
\hat{S}_{nk}(y) \equiv S_y^k(\hat{u}_{n,k}(y)), \tag{15}
\]

where we added the superscript \( k \) to \( S_y \) to denote the similarity function computed for reference frame \( k \). We set:

\[
W_{nk}(y) = \exp \left( \gamma \hat{S}_{nk}(y) \right) / \sum_{k \in \mathcal{R}_n} \exp \left( \gamma \hat{S}_{nk}(y) \right), \tag{16}
\]

where we set \( \gamma = 50 \). Eqs. (14) - (16) ensure that each row takes its recovered shift from the reference frames that exhibited good similarity. When most reference frames contribute good recoveries, Eq. (14) has the additional benefit of reducing the noise of the recovered signal by averaging.

### 4. Prototype and implementation details

Fig. 6 shows our prototype. The system consists of a rolling- and a global-shutter camera which image the same scene through a set of relay, objective and cylindrical lenses.

---

\(^4\) For example, choosing a 0.1 pixel resolution with a \( \pm 40 \) span yields the set \( V = \{(-40,-40), (-40,-39.9), (-39.9,-40), \ldots, (40,40)\} \).
Figure 5. Using multiple reference frames for recovery. (a) In this illustration, the speckle pattern’s 2D image motion follows the white arrow. During this motion, our system captures three global-shutter frames (dashed rectangles) and a single rolling-shutter frame (green rectangles are individual rows). None of the reference frames contains overlap with all the rows of the rolling-shutter frame, and therefore no single frame can be used to recover the shifts for all rows. (b) Shows the x-axis shift recovered separately using each of the reference frames. Observe that rolling-shutter rows that contain little-to-no overlap with the reference frames yield noisy recoveries. (c) Our method merges well-recovered signal portions from the multiple recoveries.

The scene is illuminated by a 532nm 4.5mW laser in a coaxial configuration using a beam-splitter seen in the left of Fig. 6. The utilized laser has relatively low power – that of standard laser pointers widely used in classrooms. Therefore, unless stated otherwise, we boost the signal in all experiments by attaching a small patch of retro-reflective tape on the surfaces we seek to measure. To capture multiple points, we spread the laser into dots by placing a diffraction grating in the cage between the laser and beam-splitter. Please see the supplementary material for a full parts list.

In Section 2, we assumed that both cameras have identical resolutions and are optically aligned. In practice, the captured images differ due to both geometric (e.g., homography, horizontal flip) and radiometric distortions. In fact, both cameras need not have the same sensor resolution. In our prototype, the rolling- and global-shutter images have resolutions of 1280x944 and 2056x1542, operating at 64.7 FPS and 134 FPS, respectively. Therefore, we need to accurately calibrate the mapping between both sensors.

Calibration includes capturing a static speckle scene, detecting and matching feature points in both frames, and computing the parameters of the desired mapping model. We used a 3rd-degree smooth bi-variate spline interpolation to compute the mapping. The mapping was computed locally per each cropped laser-point speckle column. Please see the supplementary material for more details.

We set an 18us exposure time in both cameras (unless stated differently). We reduce the rolling-shutter camera’s region-of-interest (ROI) by 40 pixels on the top and bottom of the frame so that the horizontal field-of-view of the reference camera is slightly larger than that of the rolling-shutter one. This prevents the first and last rows from shifting outside the field of view captured by the reference camera. Finally, we set hyperparameters $P = 15$ reference frames and $\lambda = (1000, 100)$ for the coarse and fine levels in Eq. (11), which yields a run time of 6 sec per frame.

5. Experimental evaluation

We demonstrate our method by capturing and replaying vibration caused by audio sources (e.g., speakers, human voice), analyzing the vibrational modes of a tuning fork, and capturing the vibrations of musical instruments. For reference, we use a high-fidelity microphone to simultaneously record the resulting sound in most experiments. Please examine supplementary material to hear the recordings [1].

5.1. Capturing audio signals

In Fig. 7 we point the camera at the membranes of two speakers. First, (Fig. 7 top row), we record one speaker playing a series of tones using a single un-split laser point. Examining the microphone and our system’s recording, one might notice that the microphone could not ‘pick up’ the low frequencies (65Hz and 33Hz). This is because a typical microphones’ frequency response is less sensitive at the lower frequencies [19]. Fig. 7(d) shows the Lissajous curves for three of the eight played tones. Observe that the speaker membrane vibrates differently between the three tones, suggesting that the three frequencies create different membrane vibration modes. In the bottom row of Fig. 7, we split the laser using a diffraction grating and measure a point on both speaker membranes simultaneously. Here, the left and right speakers are playing reversed chirp signals (up-chirp and down-chirp). While the microphone records a mixed signal (Fig. 7(g)), our system measures each speaker separately, yielding unmixed recordings (Fig. 7(h-i)).

In Fig. 8(a) we recreate an experiment similar to Davis et al. [13]. We point the system at a chips bag (with a retro-reflective patch) and play the audio file used by Davis et al. Despite the different setups, which makes a direct and fair comparison difficult, it is evident both auditorily and from the spectrograms that our system recovered the original audio with higher fidelity. The intelligibility [33] and PESQ [24] scores were [0.44, 0.73] and [1.06, 1.14], re-
Figure 7. Speaker membrane experiments. Top row: A single speaker is playing eight octaves of the note C ($C_1$ to $C_8$). The speaker’s vibrations are simultaneously recorded using our system (using a single laser point) and a microphone. (a) A spectrogram of the input signal sent to the speaker. (b) Microphone spectrogram. (c) A spectrogram of the x-axis speckle shifts recovered by our system. Observe the difference in frequency response between (b) and (c): namely, the microphone is tuned for high-frequency sensitivity, while our system has a mostly flat response. (d) Lissajous curves of three different notes show that the speaker membrane has different physical vibrations at different frequencies. Bottom row: (e-f) The left and right speakers are playing down-chirp and up-chirp signals, respectively. (g) The microphone records a mixture of both speaker signals and is unable to tell them apart. (h-i) Our system records each speaker separately (x-axis shifts shown here) and thus is able replay each channel individually.

spectively, where ours is the second score (higher is better). In Fig. 8(b), we compare to a 1D speckle sensing system using the method of Wu et al. [40] for signal recovery. The 1D system consists of a high-end line sensor [34], mounted with the same laser and objective lens used in our prototype. Using both systems, we simultaneously capture a speaker mounted on a rotating stage and playing a 523Hz tone. The played tone’s Lissajous curve (green curve in Fig. 7(d)) suggests that the speckle motion is predominantly along a single direction. Thus, rotating the speaker allows us to test the robustness of signal recovery with respect to different speckle shifts directions.

Fig. 8(b) shows that 1D speckle sensing performance strongly depends on the direction of the 2D speckle shifts. Fig. 8(b)’s left subplot shows that 1D sensing works well when the speckle motion is mostly aligned with the line sensor’s direction (x-axis). However, when a non-negligible y-axis motion component exists, the recovery breaks (Fig. 8(b) middle plot). Conversely, for the same speaker orientation, our method recovers motion in both axes correctly (Fig. 8(b) right plot).

5.2. Remote recording of musical instruments

We use our system to record musical instruments remotely. In Fig. 9(top), we record a violin played by a musician. We recover the speckle pattern’s macro-motion spanning thousands of pixels and the audio vibrations spanning only several pixels. Note that our approach could handle the natural motions of the instrument as it is being played.

In Fig. 9(bottom), we record two musicians simultaneously playing different scales on acoustic guitars. The laser beam is split to illuminate both guitars. The microphone records both guitars yielding an unpleasant dissonant recording, while our system records each guitar separately.

5.3. Analyzing vibrational modes of a tuning fork

We analyzed the vibrational modes of a 426Hz tuning fork. As shown in Fig. 10, we strike the fork with a mallet and measure its 2D vibrations at three points placed along the fork’s arm. The generated vibrational modes depend on the strike’s strength, striking position along the fork, mallet tip material, and how the fork is held [32]. While a microphone can detect various acoustic frequencies produced by the various modes, it does not provide any information on the modes’ type of motion. Conversely, by plotting the simultaneous vibration of multiple points, we can determine the kind of motion generating the mode. The measured frequencies are verified using a microphone (Fig. 10(d)).

In Fig. 10(a), the fork is struck near its head (near point 2) with a rubber-tipped mallet, mainly exciting its ‘fundamental’ mode. The fork arm’s motion can be observed by the vibration along the x-axis, whose amplitude increases from point 0 to point 2. Striking the fork with a metal bar (Fig. 10(b)), additionally induces the ‘clang’ mode, which is about $6.26 \times$ higher than the fundamental [25]. The clang mode vibrations are visible in the x-axis as a high-frequency modulating the fundamental mode. The clang mode induces an opposite phase between points 0 and 2 since these surface points tilt in opposite directions, while point 1 is approximately stationary. Hitting the fork harder (Fig. 10(c))
Figure 8. Comparisons with related methods. (a) We attempt to reproduce the chips bag experiment of Davis et al. [13]. We vibrate a chips bag using a nearby speaker and recover the speaker’s input audio. A small retro-reflective marker is attached to the chips bag for better laser reflectivity. We play the “Marry had a little lamb...” audio file used by Davis et al., and compare the resulting spectrograms between our method and the recovered audio obtained by Davis et al. Observe that our method recovers the audio with significantly improved fidelity (see supplementary material for audio comparison). Note that Davis et al. used a high-speed camera with a zoom lens and illuminated the chips bag with a strong light source. In comparison, we use a low-power laser, a marker and our low-speed dual-shutter system. (b) Comparison to 1D speckle sensing for a speaker tone [41]. Left: 1D sensing works well for x-axis only motion. Middle: But, for shifts having a non-negligible y-axis component, the reconstruction breaks. Right: We recover shifts in 2D, yielding correct waveforms.

Figure 9. Capturing vibrations of instruments. Top row: We point the system at a musician playing a violin. The chromagrams of the microphone recording and the recovered audio are shown. The recovered audio is taken as the y-axis shifts for all experiments. The right-most plot shows the recovered y-axis shifts, wherein the instruments’ musical vibrations are tiny compared to the instruments’ global motion. Bottom row: Here, we record two musicians playing two guitars simultaneously. The microphone records a mixture of both instruments, while our system records each instrument separately and is unaffected by the other instruments’ sound.

yields an additional high-frequency in-plane mode as seen in point 1’s x-axis vibrations. The sensing point positions shown in Fig. 10(top) allow the measuring of the fork’s in-plane vibrations. Out-of-plane vibrations can be analyzed by measuring the fork’s arm from above.

5.4. Sensitivity study

We examined our system’s sensitivity to tilts and transversal motion by shining a laser spot at an optomechanical stage, capable of precision tilts and transversal shifts. We measured a linear relationship between small tilts and transversal motions to the shifts of the imaged speckle pattern. The measured sensitivity to tilts was 950 and 1475 pixels/degree for the x- and y-axis, respectively, while the sensitivity to the transversal motion was 43 and 61 pixels/mm for the x- and y-axis, respectively. The $1.5\times$ factor difference in sensitivity between the axes stems from the higher optical magnification resulting from the cylindrical lens. See supplementary for plots.

6. Limitations and societal impact

Light efficiency. The performance of our method depends on the amount of light reflected from each surface point. Our prototype used a low-power laser and we enhanced light efficiency using retro-reflective tape. The supplementary material shows additional experiments that test our sys-
tem without the retro-reflective markers. A higher-power laser [5, 42] can make the tape unnecessary, however, eye safety issues may arise for applications involving humans.

**Dense vs. sparse.** Our method trades off spatial resolution for temporal resolution by sampling each time measurement using a single sensor row instead of the full sensor. While this makes our method more bandwidth efficient, a disadvantage is the inability to recover dense motion fields [13].

**Number of simultaneous measurements:** As seen in Fig. 1, each measured point occupies a horizontal fraction of the sensor plane. Thus, the number of simultaneous points is limited by the sensor’s width. As analyzed in the supplementary material, narrowing the speckle column per point (e.g. by changing the camera focus) increases the number of possible points. But, as the columns get narrower, the number of pixels per-point per-row decreases, potentially reducing correlation accuracy. Moreover, the vertical speckle makes some sensing arrangements infeasible (e.g. two points on the same image column).

**Handling object macro-motions:** We demonstrated two types of applications: recording audio (e.g. guitar, chips bag, speaker) and sensing vibrational modes (tuning fork). For the first type, our system is robust to macro motions as long as at least one laser point hits the vibrating surface. For example, for a typical guitar, one laser point allows a lateral movement range of about 13cm (half its narrowest dimension). The second type requires sensing particular points of interest, and here, object motion is unsupported by our prototype. Future methods could use a galvo-mirror to track and maintain the laser spot on the particular object point. Our method also assumes that the reference camera is fast enough to capture the speckle’s macro-motion. But fast and large object motions may yield no speckle overlay between consecutive reference frames, degrading performance.

**Rolling-shutter dead-time:** Rolling-shutter sensors typically exhibit a slight delay between the last row of a frame and the first row of the subsequent frame. This ‘dead-time’ has an insignificant effect when sensing low-frequency vibrations but may introduce noise at higher frequencies with wavelength comparable to the dead-time’s duration (0.5ms in our camera). Using our method with a line-sensor instead of a rolling-shutter can prevent the dead-time [5, 27].

**Societal impact:** Optically detecting vibrations can be useful in many scientific and engineering fields but could potentially create privacy concerns. For example, laser microphones were used to eavesdrop on distant conversations.

7. Conclusion

We present a new bandwidth-efficient approach for high-speed visual vibration sensing using two low-speed cameras. Our system can handle non-static objects and sense multiple simultaneous points (e.g., multiple musical instruments). Future works may improve signal quality using learning-based signal recovery that imposes learned priors on speech and music signals. We envision possible future applications like remotely and individually recording all instruments of an orchestra or monitoring many factory-floor machinery using just one static camera.

**Acknowledgements:** This work was supported in parts by NSF Grants IIS-1900821 and CCF-1730147. We thank A. Sankaranarayanan for advice on building the optical system, J. Smerd for playing the guitar and violin, and T. Zhang for help with the 1D speckle sensing experiments.
References


