Dual-Shutter Optical Vibration Sensing: Supplementary Materials

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

1. Additional prototype details

1.1. Hardware components

The rolling- and global-shutter cameras are IDS UI3240-NIR-GL and UI3070CP-M-GL, respectively. The relay optics consists of a beam splitter (Thorlabs CCM1-BS013) and a 75mm matched doublet pair (Thorlabs MAP1075150-A). The objective and cylindrical lenses are Edmund optics 16mm/f1.8 and Thorlabs 50mm LJ4709RM-A, respectively. The cylindrical lens is placed approximately 25mm from the objective lens. The green laser is a 532nm 4.5mW (Thorlabs CPS532) laser. The diffraction gratings we used to spread the laser into multiple dots are Edmund optics 80 and 92 grooves/mm.

1.2. Creating multiple laser points

In Figs. 1,7,9,10 of the main paper, we captured multiple laser points simultaneously. Here, we provide more details on how we generated these multiple points. Fig 1(a) shows a schematic of the optics used to generate multiple scene points. The laser beam passes through a diffraction grating, splitting the beam into multiple beams exiting at different angles. The number of beams depends on the diffraction grating type. The outgoing beams are relayed towards the scene through a beam-splitter (Thorlabs PBSW-532R) to illuminate the object (or objects) points of interest. The dual-shutter camera views the scene through the beam splitter in a coaxial configuration to maximize the signal intensity.[1]

The multiple outgoing laser beams illuminate several points of interest on the object surface. In a focused system without the cylindrical lens, the illuminated points project to a set of corresponding image-plane points via standard projective geometry (Fig 1(b)). Defocusing the objective lens spreads each image point into a circle in which the speckle pattern is visible (Fig 1(c)). Finally, the cylindrical lens vertically spreads each circle into a (speckle) column (Fig 1(d)).

In Fig 1(a), the laser beams exiting to the right of the beam splitter (exit 2) are blocked. However, these beams can used instead of the forward facing beams (exit 1) shown in the schematic. For example, in the two-guitar experiment shown in Fig. 9(bottom) of the main paper, only beam 2 was unblocked at exit 1 of the beam splitter, while only beam 4 was unblocked at exit 2. Note that each split beam can only be used at one of the beam splitter exits, and must be blocked on the other.

[1]The coaxial configuration is mostly useful when using retro-reflective markers. Without any markers, the laser can illuminate the surface directly without the beam-splitter.

2. Coarse-to-fine 2D speckle shift recovery

Eqs. (11)-(13) of the main paper describe a method for computing the per-row shifts $\delta\hat{\mathbf{u}}_{nk}(y)$. This method can be applied *as is* for small to moderate dictionaries $\mathcal{V}$. However, as the number of possible shifts M increases, computing the similarity measure for all possible shifts becomes a computationally expensive operation. Therefore, we provide a two-step coarse-to-fine approach for recovering $\delta\hat{\mathbf{u}}_{nk}(y)$. We experimentally tested the proposed approach and found it gives sufficiently identical results to the direct approach described by Eqs. (11)-(13), for a fraction of the run time.

The solution is split into two levels: coarse and fine. Define

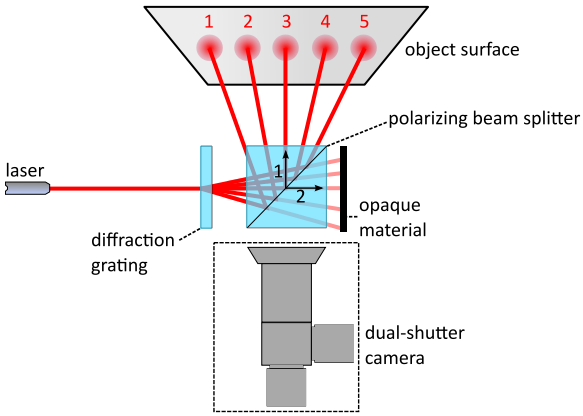
$$\delta\hat{\mathbf{u}}_{nk}(y) \equiv \delta\hat{\mathbf{u}}^{\mathrm{c}}(y) + \delta\hat{\mathbf{u}}^{\mathrm{f}}(y), \tag{1}$$

where $\delta\hat{\mathbf{u}}^{\mathrm{c}}(y)$ and $\delta\hat{\mathbf{u}}^{\mathrm{f}}(y)$ are the coarse and fine components, respectively. Here we dropped the nk subscript for brevity. In the coarse level, the shifts $\delta\hat{\mathbf{u}}^{\mathrm{c}}(y) = \left(\delta\hat{u}^{\mathrm{c}}_{dx}(y), \delta\hat{u}^{\mathrm{c}}_{dy}(y)\right)$ are recovered up to a single pixel resolution, while the fine level refines the solution for a sub-pixel resolution.

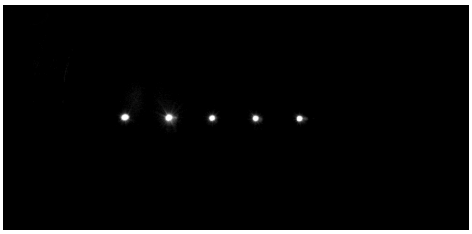
Coarse level: In the coarse level, the shifts are recovered sequentially: first we recover the y-axis shifts, followed by the x-axis shifts. We normalize the rows of $\bar{I}_{\mathrm{RS}}$ and $\bar{I}_{\mathrm{GS}}$ yielding $\bar{I}^{\mathrm{norm}}_{\mathrm{RS}}$ and $\bar{I}^{\mathrm{norm}}_{\mathrm{GS}}$, respectively. This row-wise normalization consists of subtracting the mean of each row and dividing each row by the row's standard deviation. Then we apply a row-wise Fast Fourier Transform on the rows of $\bar{I}^{\mathrm{norm}}_{\mathrm{RS}}$ and $\bar{I}^{\mathrm{norm}}_{\mathrm{GS}}$, yielding $\bar{I}^{\mathcal{F}}_{\mathrm{RS}}$ and $\bar{I}^{\mathcal{F}}_{\mathrm{GS}}$ respectively.[2]

Let $\mathcal{Y} = \{y_l\}_{l=0}^{L-1}$ denote a set of L possible y-axis shifts having some maximum span, and a step size of one pixel,

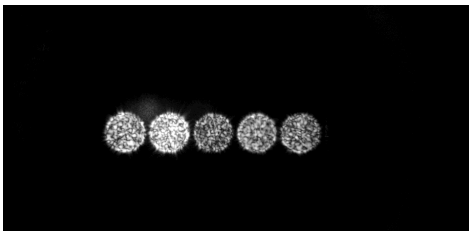
[2]Both $\bar{I}^{\mathcal{F}}_{\mathrm{RS}}$ and $\bar{I}^{\mathcal{F}}_{\mathrm{GS}}$ are shifted using the $\mathrm{fftshift}()$ function such that the zero-frequency component is at the center of the array.



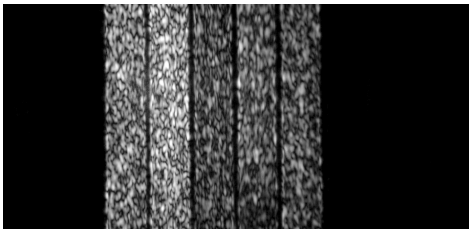
(a) generating multiple laser points



(b) focusing on the object surface



(c) defocusing the objective lens



(d) adding the cylindrical lens

Figure 1. Generating multiple laser points. **(a)** We generate multiple laser points by passing the laser through a diffraction grating. **(b)** Without the cylindrical lens, focusing the objective lens at the object surface creates a "standard" image of the laser points. **(c)** Defocusing the objective lens, turns each laser points into a circle in which the speckle pattern is visible. **(d)** Adding the cylindrical lens in front of the objective lens spreads each circle into a (speckle) column.

e.g. $\mathcal{Y} = \{-40, -39, .., 40\}$. Next we compute the correlation for the normalized rows in the Fourier domain [6] for each shift in $\mathcal{Y}$:

$$O_y(x, y_l) = \frac{1}{R} \left| \mathcal{F}^{-1} \left(\bar{I}^{\mathcal{F}}_{\mathrm{RS}}(\mathbf{x}, t^{\mathrm{rs}}_n) \odot \bar{I}^{\mathcal{F}}_{\mathrm{GS}}(\mathbf{x} + (0, y_l), t^{\mathrm{gs}}_k)^{\mathrm{conj}} \right) \right|, \tag{2}$$

where $\mathcal{F}^{-1}(\cdot)$ is the inverse FFT operator, R is the width of the speckle column in pixels, and the superscript conj denotes a complex conjugate. Notice that for every vertical shift y_l, Eq. (2) yields a vector of normalized correlations for R horizontal shifts in the range $\{-R/2, .., R/2\}$. Therefore, function $O_y(x, y_l)$ simultaneously provides information on both x- and y-axis correlations between the rolling- and global-shutter frames. Specifically, we expect that for the correct vertical shift y_l, the peak correlation value across all x-axis shifts in $O_y(x, y_l)$ will be the highest with respect to other y_l. Moreover, once the correct vertical shift y_l is found, the location of the highest correlation peak directly corresponds to the recovered x-axis shift.

Therefore, we first compute the optimal vertical y-axis shift. Let

$$S^{dy}_y(y_l) = \max_x O_y(x, y_l) \tag{3}$$

denote the y-axis similarity measure. Define $\mathcal{U}^{dy} = \{\delta\hat{u}^{\mathrm{c}}_{dy}(y)\}_{\forall y}$ as the set of y-axis shifts for all rows, where $\delta\hat{u}^{\mathrm{c}}_{dy}(y) \in \mathcal{Y}$.

We recover the y-axis shift by minimizing the following loss function:

$$E(\mathcal{U}^{dy}) = \sum_y \left[1 - S^{dy}_y(\delta u^{\mathrm{c}}_{dy}(y))\right] + \lambda \sum_{y,y'} V_{y,y'}(\delta u^{\mathrm{c}}_{dy}(y), \delta u^{\mathrm{c}}_{dy}(y')), \tag{4}$$

where the solution is given by

$$\hat{\mathcal{U}}^{dy} = \underset{\mathcal{U}^{dy}}{\operatorname{argmin}} \left(E(\mathcal{U}^{dy}) \right), \tag{5}$$

and all other terms in Eq. (4) are analogous to the same terms in Eq. (11) in the main paper. Next we compute the x-axis shifts using the recovered y-axis shifts.

Let $\mathcal{U}^{dx} = \{\delta\hat{u}^{\mathrm{c}}_{dx}(y)\}_{\forall y}$ denote the set of all x-shifts, where $\delta\hat{u}_{dx}(y) \in \mathcal{X}$, and $\mathcal{X} = \{x_r\}_{r=0}^{R-1}$. The similarity measure for the x-axis shifts is given by

$$S^{dx}_y(x_r) = O_y\left(x_r, \delta\hat{u}^{\mathrm{c}}_{dy}(y)\right). \tag{6}$$

Then we recover $\hat{\mathcal{U}}^{dx}$ by minimizing the loss function:

$$E(\mathcal{U}^{dx}) = \sum_y \left[1 - S^{dx}_y(\delta\hat{u}^{\mathrm{c}}_{dx}(y))\right] + \lambda \sum_{y,y'} V_{y,y'}(\delta\hat{u}^{\mathrm{c}}_{dx}(y), \delta\hat{u}^{\mathrm{c}}_{dx}(y')), \tag{7}$$

using

$$\hat{\mathcal{U}}^{dx} = \underset{\mathcal{U}^{dx}}{\operatorname{argmin}} \left(E(\mathcal{U}^{dx}) \right). \tag{8}$$

Fine level: After recovering the coarse level shifts, the fine level shifts $\delta\hat{\mathbf{u}}^{\mathrm{f}}(y)$ are recovered using the standard procedure described in the main paper except for slight modifications. The set of fine level shifts $\mathcal{V}$ is now set to a sub-pixel resolution (*e.g.* $\mathcal{V} = \{(-0.5, -0.5), (-0.5, -0.4), .., (0.5, 0.5)\}$. Recovery is done using Eqs. (11), (13) of the main paper, along with

an augmented Eqs. (12) that accounts for the coarse level shifts:

$$S_y(\mathbf{v}_m) = \mathrm{ZNCC}\left(\bar{I}_{\mathrm{RS}}(\mathbf{x}, t_n^{\mathrm{rs}}), \bar{I}_{\mathrm{GS}}(\mathbf{x} + \delta\hat{\mathbf{u}}^{\mathrm{c}}(y) + \mathbf{v}_m, t_k^{\mathrm{gs}})\right). \tag{9}$$

The recovered shifts $\delta\hat{\mathbf{u}}^{\mathrm{f}}(y)$ are added to the coarse level shifts using Eq. (1) to yield the final result.

3. Reference frame selection

As shown in Fig. 5 of the main paper, high-amplitude motions may cause a single reference frame to be insufficient for recovering all row shifts in frame n. Therefore, using multiple frames improves signal recovery by increasing the chance that all rolling-shutter rows in frame n will have a corresponding overlap in one of the reference frames. A simple solution would be to use *all* reference frames for recovering each rolling-shutter frame; however, this would yield much longer run times. Instead, we limit recovery to a set of P reference frames, which must be selected for each frame n, as explained below.

We use two strategies of reference frame selection which depend on the object's macro-motion. In non-static objects, such as hand-held musical instruments, the low-frequency motion amplitude may be substantial, spanning thousands of pixels in both axes. Therefore, it is highly likely that only temporally adjacent global-shutter frames will contain any overlap with any given rolling-shutter frame n. Consequently, in scenes with large motions, we set $\mathcal{R}_n$ to the P frames whose timestamp t_k^{gs} is closest to t_n^{rs}.

In scenes where the object's low-frequency motion amplitude is low (*e.g.* chips bag, tuning fork, or speaker membrane), the global speckle pattern drift across time is relatively small, spanning just a few dozen pixels. Therefore, for each rolling-shutter frame n, the set of 'relevant' reference frames which may have significant overlap with frame n is larger than in the non-static case. Two reference frames k_1 and k_2 whose global shift is nearly identical $\mathbf{u}(t_{k_1}^{\mathrm{gs}}) \approx \mathbf{u}(t_{k_2}^{\mathrm{gs}})$ will likely have similar overlaps with frame n, and thus will contribute redundant information. Instead, we wish to select P frames that provide the largest cover of the 2D speckle pattern.

Let $\tilde{\mathcal{R}}_n = \{k_0, k_1, .., k_{Q-1}\}$ denote a set of Q temporally closet reference frames to frame n, where $Q > P$. Without loss of generality, suppose that indices in $\tilde{\mathcal{R}}_n$ are ordered by the proximity of the frame's timestamp to t_n^{rs}, namely that k_0 belongs to the reference frame whose timestamp $t_{k_0}^{\mathrm{gs}}$ is closest to t_n^{rs}. We select the P reference frames from $\tilde{\mathcal{R}}_n$ as described in Algorithm 1. The main idea here is to iteratively select reference frames whose global shift is the farthest away from all the shifts in the selected reference frames so far. We used $P = 15$ for all scenes and $Q = 30$ for static scenes.

Algorithm 1 Selecting P reference frames from $\tilde{\mathcal{R}}_n$

Initialize $\mathcal{R}_n \longleftarrow \emptyset$.
Initialize $\tilde{\mathcal{R}}_n \longleftarrow \{k_0, k_1, .., k_{Q-1}\}$.
Add closest frame index to $\mathcal{R}_n \longleftarrow \mathcal{R}_n \cup k_0$
Subtract closest frame index from $\tilde{\mathcal{R}}_n \longleftarrow \tilde{\mathcal{R}}_n \setminus k_0$
while $||\mathcal{R}_n|| < P$ **do**
    Find index $k_j \in \tilde{\mathcal{R}}_n$ that is farthest from the set $\mathcal{R}_n$,
    $k_j = \underset{k_j \in \tilde{\mathcal{R}}_n}{\mathrm{argmax}} \ \underset{k_i \in \mathcal{R}_n}{\min} \left(|\mathbf{u}(t_{k_i}^{\mathrm{gs}}) - \mathbf{u}(t_{k_j}^{\mathrm{gs}})|\right)$
    Add k_j to $\mathcal{R}_n \longleftarrow \mathcal{R}_n \cup k_j$
    Subtract k_j from $\tilde{\mathcal{R}}_n \longleftarrow \mathcal{R}_n \setminus k_j$
end while

4. Global-shutter to rolling-shutter mapping

In section 4 of the main paper, we outline the calibration process of computing a mapping function between the global and rolling shutter image domains. In this section, we provide a more detailed description of this process.

The calibration process consists of three main stages:
(a) We capture and store a pair of frames of a static object. Then, we extract feature points in both frames using a SIFT descriptor [5]. Using the extracted feature points, we estimate an initial 'rough' homography transform between the full sensor frames. Let I_{RS}^0 and I_{GS}^0 denote the stored calibration frames. The initial homography mapping is insufficiently accurate since it can not encapsulate non-projective lens distortions. However, it will be used to automatically find and crop roughly the same image domain in the global-shutter frames.

During vibration sensing, we point the system at the object or objects of interest and record the simultaneous videos. We crop the rolling-shutter video I_{RS} on the speckle column of the point we wish to recover, yielding $\bar{I}_{\mathrm{RS}}$.
(b) We then apply the same crop on I_{RS}^0 yielding $\bar{I}_{\mathrm{RS}}^0$. The initial homography is then used to automatically find and crop the same image domain in the global-shutter frames for the calibration frame and the captured vibration video, yielding $\bar{I}_{\mathrm{GS}}^0$ and $\bar{I}_{\mathrm{GS}}$, respectively.
(c) Next, we compute a more accurate mapping by repeating the feature extraction process on the cropped $\bar{I}_{\mathrm{RS}}^0$ and $\bar{I}_{\mathrm{GS}}^0$, and using the extracted points to fit a 3rd-degree smooth bivariate spline interpolation between the frames. Finally, we apply the resulting mapping to I_{GS} to yield the desired $\bar{I}_{\mathrm{GS}}$.

Stage (a) is only done once, while stages (b) and (c) are repeated before the recovery of every unique speckle column to yield the most accurate local mapping. [3] This calibration process takes a few seconds to complete.

Imperfections in the resulting mapping may still yield a small sub-pixel bias to the x- and y-shifts, respectively.

[3] A unique speckle column is a speckle column whose column range in I_{RS} has not been calibrated yet.

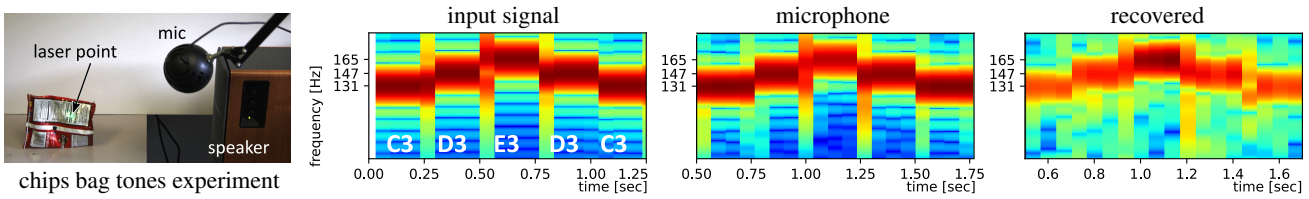


Figure 2. Chips bag experiment without markers. We image the chips bag of Fig. 8(a) of the main paper without any retro-reflective markers. The speaker is playing a series of tones which are simultaneously recorded using our system and a microphone. The spectrograms of the input signal, microphone recording, and recovered audio are shown.

We remove these biases by averaging the signal of a few static frames (where the object is static), and subtracting the computed x- and y- axis biases, per frame, from all future measurements.

5. Post processing for audio replay

We post-process the recovered speckle shifts intended for an audio replay. First, we apply a high-pass filter to get rid of the object's low-frequency macro motions. Then, similar to Davis *et al*. [3], we interpolate the rolling-shutter dead-time between subsequent frames using either an ad-hoc method based on Fourier interpolation or a technique based on fitting an autoregressive model [4]. Both approaches yielded nearly identical results aurally. We then identify large spikes in the measurements using a simple detection algorithm, where we threshold based on the difference between the original and a median-filtered signal. We remove the detected spikes and interpolate new values for those timestamps. Finally, we use audio editing software to crop the audio and apply an additional denoising step where we generate a noise profile using a short interval of silence and subtract the resulting profile in FFT-domain [2]. In the last denoising step, the noise profile is generated once and applied to recoveries from the same experiment.

6. Additional experimental results

6.1. Chips bag without retro-reflective marker

Fig. 8(a) of the main paper shows an experiment where our system captures and replays the indirect vibrations of a chips bag reacting to audio from a nearby speaker. In Fig. 8(a) the chips bag is attached with a small retro-reflective marker to boost the light reflected from the low-power laser used in our prototype. Fig. 2 shows an experiment where we capture the bag's vibration without any markers.

To compensate for the low SNR, we increase the camera's exposure time to 2.5ms, corresponding to an effective temporal sampling rate of 400Hz. We record the nearby speaker playing a series of tones, and show the spectrograms for the input signal, microphone, and the signal recovered using our system. Note that our prototype utilized a low-power laser (4.5mW). A higher power laser as in [1,7] can enable capturing audio at higher frequencies, similar to Fig. 8(a), without any markers.

6.2. Sensitivity study and macro-motion velocity

In Section 5.4 of the main paper, we describe a sensitivity study in which we measure the relationship between small tilts and transversal motions of the observed surface to 2D speckle pattern shifts in the image plane. Fig. 3 shows the resulting linear relationships. As stated in the main paper, the measured sensitivity to tilts was 950 and 1475 pixels/degree for the x- and y-axis, respectively, while the sensitively to the transversal motion was 43 and 61 pixels/mm for x- and y-axis, respectively. The $\sim 1.5\times$ factor difference in sensitivity between the axes stems from the higher optical magnification resulting from the cylindrical lens.

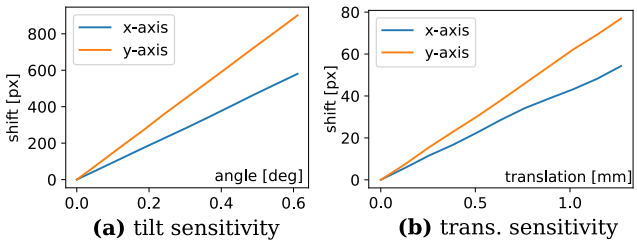
The sensitivity figures above can be useful to assess the maximum physical object velocities that the system can handle.[4] Object marco-motions lead to speckle pattern shifts in the image plane. Our method assumes that the reference camera is fast enough (or that the object motion is slow enough), such that there exists some minimum overlap between consecutive reference frames.

The maximum allowable velocity depends on many factors (*e.g*. object distance, speckle column width, SNR, optics). Assuming a 200px-wide column and 10% minimum overlap between frames, the speckle can shift up to 180 pixels per frame. Since the camera operates at 134 FPS, this means that the pattern can shift up to $180\times134=24120$ pixels per second. Dividing the 24120 pixels-per-second by the x-axis sensitivity of 43 pixels/mm, yields a maximum physical transversal velocity of 560mm/sec or about 0.6m/sec. Note that this velocity assessment holds for the object distance at which the sensitivity was measured.

6.3. Correlation accuracy vs. pixels per row

In Section 6 of the main paper, we asserted that as the speckle columns get narrower, the number of pixels per-point per-row decreases, which can potentially decrease the

[4]The object's motion in space can be handled as long as the laser hits the object's vibrating surface.



(a) tilt sensitivity **(b)** trans. sensitivity

Figure 3. Sensitivity study. We point the system at a optomechanical stage capable of precise tilts, shift and rotation, and record the resulting 2D speckle image motion. The sensitivity to tilts was 950 and 1475 pixels/degree for x- and y-axis, respectively. Shift sensitivity is 43 and 61 pixels/mm for x- and y-axis, respectively. The camera was placed approximately 75mm away from the stage.

correlation accuracy (Eq. (12) of the main paper). In general, besides the pixels per row, the correlation accuracy also depends on several other factors such as the x-axis shift amplitude, the SNR and more. In this experiment, we analyze the correlation accuracy as a function of the speckle column width while keeping the SNR constant and minimizing the effect of x-axis shift amplitude by observing a signal with mostly y-axis shifts.

Fig. 4 shows the mean squared error (MSE) for signal recovery as a function of the speckle column width in pixels. The error is computed with respect to the "nominal" speckle column width of 135 pixels. The plot shows no significant increase in error up to around 35px-wide columns (red point in Fig. 4). When decreasing the column width below 35 px, the accuracy rapidly drops.

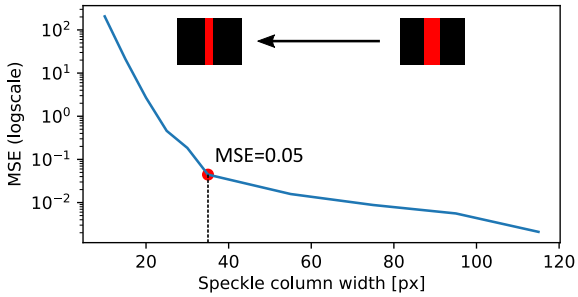


Figure 4. Correlation accuracy vs. pixels per row. Each laser point creates a 'speckle column' in the image plane. Decreasing the width of the speckle column in pixels (x-axis in the plot) per point may decrease the correlation accuracy. The plot shows that decreasing the speckle column width has negligible effect up to a width of 35 pixels, after which the reconstruction mean squared error (MSE) begins to deteriorate.

References

[1] S Bianchi and E Giacomozzi. Long-range detection of acoustic vibrations by speckle tracking. *Applied optics*, 58(28):7805–7809, 2019. 4

[2] Cockos Incorporated. REAPER. https://www.reaper.fm/. 4

[3] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. The visual microphone: Passive recovery of sound from video. 2014. 4

[4] AJEM Janssen, R Veldhuis, and L Vries. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(2):317–330, 1986. 4

[5] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 3

[6] Wikipedia contributors. Cross-correlation, 2004. [Online; accessed 23-November-2021]. 2

[7] Zeev Zalevsky, Yevgeny Beiderman, Israel Margalit, Shimshon Gingold, Mina Teicher, Vicente Mico, and Javier Garcia. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Optics express*, 17(24):21566–21580, 2009. 4