2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897338

COMPARISON OF PHASE-BASED SUB-PIXEL MOTION ESTIMATION METHODS

Cédric Marinel[⋆†] *Benjamin Mathon*[⋆] *Olivier Losson*[⋆] *Ludovic Macaire*[⋆]

[⋆] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[†] EOMYS Engineering, Lille, France
Email: cedric.marinel@eomys.com

ABSTRACT

Monitoring mechanical properties of structures such as bridges, buildings, or wind turbines, is important to detect early stage failures. Operational modal analysis is a testing method to estimate these properties from vibration measurements. Taking advantage of works on motion estimation, several video modal analysis methods have emerged in the last decade. This paper compares two strategies about sub-pixel phase-based motion estimations thanks to multi-scale decomposition. Synthetic videos of a vibrating cantilever beam are generated to assess the robustness of these methods against motion amplitude, gray scale quantization, white noise, and blurring.

***Index Terms*—** phase-based motion estimation, sub-pixel motion, multi-scale decomposition, vibration, mechanical structure

1. INTRODUCTION

As mechanical properties reflect the health of civil structures, their analysis helps to detect early stage failure. Operational modal analysis (OMA) has been developed to estimate structure mechanical properties, such as mode shape and natural frequencies, from acceleration, velocity, or displacement measurements [1]. These measurements are generally acquired by contact sensors such as accelerometers or linear variable differential transformers. However, as these sensors are physically mounted on structures, their setup is costly and time-consuming. Therefore, video-based methods have been recently developed to perform cheap remote measurements from observations by still cameras [2–5]. As vibration can be defined as a periodic small motion of surface elements observed by the camera, motion should be estimated at sub-pixel definition. Chen et al. [6] use the phase complex response of quadrature filters applied to video frames to estimate motion and perform vibration analysis of simple structures. Chou et al. [5] experimentally show that among available classical video motion estimation methods, the phase-based one is faster and provides a dense sub-pixel motion estimation. Furthermore, phase-based methods do not require speckle patterns mounted or projected on the structure.

Two approaches exist to estimate motion thanks to multi-scale pyramid decomposition of each frame. Wadhwa et al. [7] take the pyramid scales into account to estimate motion at each pixel, so that one single OMA is performed, whereas Yang et al. [8] determine the motion at a given scale.

Because no study compares phase-based motion estimation methods, we propose to assess their performances using synthetic videos that represent a vibrating vertical cantilever beam. We then compare estimated displacements to ground-truth ones. Videos are generated with different motion amplitudes to study sub-pixel efficiency. Robustness against gray scale quantization, additive noise, and blurring is also studied.

2. PHASE-BASED MOTION ESTIMATION

2.1. Complex steerable pyramid decomposition

Let $I(x, y; t)$ be the intensity at pixel (x, y) at frame t. One wants to densely estimate the motion field along horizontal and vertical directions at each frame t:

$$\delta(x, y; t) = \begin{pmatrix} \delta^h(x, y; t) \\ \delta^v(x, y; t) \end{pmatrix} \in \mathbb{R}^2. \quad (1)$$

Assuming illumination is spatially and spectrally constant over time, the intensity at a pixel associated to a given surface element can be considered as constant:

$$I(x, y; 0) = I(x + \delta^h(x, y; t), y + \delta^v(x, y; t); t). \quad (2)$$

The following methods rely on a complex steerable pyramid (CSP) decomposition to split each frame in space frequency sub-band using quadrature complex filters. Spatial frequencies of each frame are decomposed as $(\omega^h, \omega^v) = (\omega_r \cos(\theta), \omega_r \sin(\theta))$ into polar coordinates corresponding to different scales $r = 1, \ldots, \mathcal{N}_r$ and orientations $\theta = 0, \ldots, \frac{\mathcal{N}_\theta - 1}{\mathcal{N}_\theta}\pi$, where $\mathcal{N}_r$ and $\mathcal{N}_\theta$ are the number of scales and orientations.

Magnitude and phase of the complex response $S_{r,\theta}(x, y; t) = G_{r,\theta} * I(x, y; t)$ of the filter $G_{r,\theta}$ applied to frame I are denoted as $\rho_{r,\theta}(x, y; t)$ and $\varphi_{r,\theta}(x, y; t)$.

2.2. Phase-based motion estimation

Filter response at frame 0 can be expressed from motion at frame t:

$$S_{r,\theta}(x,y;0) = G_{r,\theta} * I(x,y;0) \tag{3}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(\tilde{x},\tilde{y};0)\cdot G_{r,\theta}(x-\tilde{x},y-\tilde{y})\,d\tilde{x}\,d\tilde{y} \tag{4}$$

$$\overset{(2)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(\tilde{x}+\delta^h(\tilde{x},\tilde{y};t),y+\delta^v(\tilde{x},\tilde{y};t);t)\cdot G_{r,\theta}(x-\tilde{x},y-\tilde{y})\,d\tilde{x}\,d\tilde{y}. \tag{5}$$

Assuming that δ^h and δ^v are constant in the support (of size $b_r^h \times b_r^v$) of the filter $G_{r,\theta}$ at each frame t, we have:

$$S_{r,\theta}(x,y;0) = \int_{x-b_r^h}^{x+b_r^h}\int_{y-b_r^v}^{y+b_r^v} I(\tilde{x}+\delta^h(t),\tilde{y}+\delta^v(t);t)\cdot G_{r,\theta}(x-\tilde{x},y-\tilde{y})\,d\tilde{x}\,d\tilde{y}. \tag{6}$$

Using the changes of variables $\hat{x} = \tilde{x}+\delta^h(t)$ and $\hat{y} = \tilde{y}+\delta^v(t)$, filter response at frame 0 can be deduced from filter response at frame t as:

$$S_{r,\theta}(x,y;0) = \int_{x-\delta^h(t)-b_r^h}^{x-\delta^h(t)+b_r^h}\int_{y-\delta^v(t)-b_r^v}^{y-\delta^v(t)+b_r^v} I(\hat{x},\hat{y};t)\cdot G_{r,\theta}(x+\delta^h(t)-\hat{x},y+\delta^v(t)-\hat{y})\,d\hat{x}\,d\hat{y} \tag{7}$$

$$= S_{r,\theta}(x+\delta^h(t),y+\delta^v(t);t). \tag{8}$$

For each sub-band, filter response magnitude and phase are thus expressed as:

$$\rho_{r,\theta}(x,y;0) = \rho_{r,\theta}(x+\delta^h(t),y+\delta^v(t);t), \tag{9}$$

$$\varphi_{r,\theta}(x,y;0) = \varphi_{r,\theta}(x+\delta^h(t),y+\delta^v(t);t). \tag{10}$$

By considering that $\delta^h(t)$ and $\delta^v(t)$ depend on pixel location, we deduce that filter response magnitude and phase at a pixel associated to a surface element are nearly constant:

$$\rho_{r,\theta}(x,y;0) \approx \rho_{r,\theta}(x+\delta^h(x,y;t),y+\delta^v(x,y;t);t), \tag{11}$$

$$\varphi_{r,\theta}(x,y;0) \approx \varphi_{r,\theta}(x+\delta^h(x,y;t),y+\delta^v(x,y;t);t). \tag{12}$$

We suppose that $\varphi_{r,\theta} \in \mathcal{C}^1$ for all r,θ, and t to apply a first-order Taylor expansion to Eq. (12):

$$\varphi_{r,\theta}(x,y;0) - \varphi_{r,\theta}(x,y;t) \approx \nabla\varphi_{r,\theta}(x,y;t)\cdot\delta(x,y;t). \tag{13}$$

Because phase gradient $\nabla\varphi_{r,\theta}$ is approximately equal to the filter central spatial frequencies [9], motion can be estimated by replacing $\nabla\varphi_{r,\theta}$ by (ω^h,ω^v) in Eq. (13). Let us use the Dirac comb to sample continuous space-time quantities I, $\rho_{r,\theta}$, and $\varphi_{r,\theta}$, and denote them in discrete space as $I[x,y;t]$, $\rho_{r,\theta}[x,y;t]$, $\varphi_{r,\theta}[x,y;t]$, and $\delta[x,y;t]$, with $[x,y;t] \in [\![1,\mathcal{N}_x]\!] \times [\![1,\mathcal{N}_y]\!] \times [\![0,\mathcal{N}_t]\!]$, where $\mathcal{N}_x$, $\mathcal{N}_y$ and $\mathcal{N}_t$ are the number of pixel columns, pixel rows, and frames.

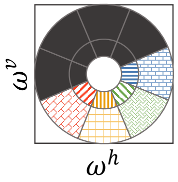

Fig. 1. Ideal Simoncelli and Freeman frequency filters $G_{r,\theta}$ support of a pyramid with $\mathcal{N}_r = 2$ scales and $\mathcal{N}_\theta = 4$ orientations.

2.3. Multi-scale motion estimation

To decompose each frame into a CSP, Wadhwa et al. [7] use Simoncelli and Freeman [10] frequency filters whose support are represented in Fig. 1. They solve a weighted least square (WLS) problem to estimate motion $\hat{\delta}$ by fusing sub-band phases:

$$\hat{\delta}[x,y;t] = \underset{\delta[x,y;t]}{\operatorname{argmin}} \sum_{r,\theta}\sum_{k=-9}^{9}\sum_{\ell=-9}^{9} \mathcal{G}[k,\ell]\cdot \rho_{r,\theta}^2[x+k,y+\ell;t]\cdot\Big[\big(\omega^h,\omega^v\big)\cdot\delta[x,y;t] - \big(\varphi_{r,\theta}[x+k,y+\ell;0]-\varphi_{r,\theta}[x+k,y+\ell,t]\big)\Big]^2. \tag{14}$$

Weights are based on the squared filter response magnitude $\rho_{r,\theta}^2$ used for sub-band decomposition. Indeed, the phase at a sub-band is meaningful only if the associated magnitude is high. The authors also assume that the motion is locally constant and add a spatial constraint. Therefore, they apply a Gaussian kernel $\mathcal{G}$ with a standard deviation of 3 px and a support of 19×19 px. Furthermore, the phase $\varphi_{r,\theta}[x,y;t]$ is wrapped in $(-\pi,\pi]$. Before solving Eq. (14), it is temporally unwrapped to compare phase shift between frames t and 0. In Eq. (14), $\varphi_{r,\theta}$ and $\rho_{r,\theta}$ are bi-cubic interpolated for $r \geq 2$ to get the same spatial resolution as $\varphi_{1,\theta}$ and $\rho_{1,\theta}$.

2.4. Single-scale horizontal motion estimation

Yang et al. [8] also use the frequency filters [10] of Fig. 1 for CSP frame decomposition. Furthermore, they assume that vertical displacement in their vertical cantilever beam videos can be neglected (i.e., $\delta^v(x,y;t) \approx 0$), such that Eq. (10) yields:

$$\varphi_{r,\theta}(x,y;0) \approx \varphi_{r,\theta}(x+\delta^h(x,y;t),y;t). \tag{15}$$

Therefore, using Taylor expansion and phase partial derivative approximation [9], Eq. (13) becomes:

$$\varphi_{r,\theta}(x,y;0) = \varphi_{r,\theta}(x,y;t) + \omega^h\cdot\delta^h(x,y;t). \tag{16}$$

The authors only use the response of horizontal filters ($\theta = 0$) and estimate horizontal motion at scale $r \in \{1,2\}$ by:

$$\hat{\delta}_r^h[x,y;t] = \frac{\varphi_{r,0}[x,y;0]-\varphi_{r,0}[x,y;t]}{\omega_r}. \tag{17}$$

Force f	0.04	0.08	0.16	0.33	0.65	1.31	2.62	5.24	10.47
δ^h at top	0.016	0.03	0.06	0.13	0.25	0.50	1.00	2.00	4.00
δ^h at middle	0.005	0.01	0.02	0.04	0.09	0.18	0.35	0.70	1.40

Table 1. Amplitude of true horizontal motion δ^h at top and middle edge pixels (px) vs. input force f (N).

Phase is also temporally unwrapped before motion estimation. $\hat{\delta}^h_1$ is computed at frame resolution whereas $\hat{\delta}^h_2$ is first computed with sub-sampled phase $\varphi_{2,0}$, then up-sampled using bi-cubic interpolation to get the frame full spatial resolution.

3. EXPERIMENTS

3.1. Experimental setup

To compare these methods, we generate synthetic videos of a vertical cantilever beam. This model requires adjusting the following physical beam parameters : length L (m), section area (m^2), Young modulus E (Pa), moment of inertia J (m^4) and mass per unit length μ (kg·m^{-1}). The center line of the vertical beam is defined in the scene coordinates system by the point set $\{(g(Y,t), Y; t) \in \mathbb{R} \times [0, L] \times [\![0, \mathcal{N}_t - 1]\!]\}$, where g is solution of the Euler-Bernoulli equation:

$$EJ\frac{\partial^4 g(Y,t)}{\partial Y^4} + \mu\frac{\partial^2 g(Y,t)}{\partial t^2} = f(Y,t). \qquad (18)$$

In this experiment, the input force f (N) is represented by a time and space Dirac function to simulate a hammer impact at the free end of the beam. Besides, to simulate the behavior of our experimental beam, we set its volume to $900 \times 30 \times 6\,\text{mm}^3$, its mass to $1.413\,\text{kg}$, and its Young modulus E to $210 \cdot 10^9$ Pa.

Frame definition is set to $720 \times 40\,\text{px}$ and since the beam covers 97% of the frame height, the pixel size is $1.289\,\text{mm}$. Frame rate is set to $436\,\text{fps}$, which fits our experimental camera. Vibrations with frequency lower than $218\,\text{Hz}$ can then be studied during $2.3\,\text{s}$ thanks to the analysis of $\mathcal{N}_t = 1000$ frames. Each pixel intensity value is computed with the area of the intersection between the pixel in the image plane and the projected beam. Values are then scaled between 30 and 225 to encode the gray level of each pixel on $\mathcal{N}_b = 8\,\text{bits}$.

We focus on two edge pixels, hereafter called top and middle edge pixels (see Fig. 2). Table 1 shows the amplitude of the true (model-based) horizontal motion δ^h computed at these pixels for a given input force f. At the top edge pixel, the amplitude of δ^h is approximately three times that at the middle edge pixel. Examples of synthetic videos of the cantilever beam can be downloaded from the following link: `https://bit.ly/3ynSEN1`.

To estimate δ^h, each video is analyzed by the multi-scale approach using Eq. (14) with $\mathcal{N}_r = 2$ and $\mathcal{N}_\theta = 4$ to obtain $\hat{\delta}^h$, and by the single scale approach using Eq. (17) to compute $\hat{\delta}^h_1$ and $\hat{\delta}^h_2$. As the thickness of the beam covers 5 px, we use a different Gaussian kernel $\mathcal{G}$ in the multi-scale approach with a standard deviation of 1 px and a support of $7 \times 7\,\text{px}$. For illustration purpose, true and estimated horizontal motions are displayed in Fig. 3 for a beam excited by an input force $f = 10.47\,\text{N}$. This figure shows that $\hat{\delta}^h$ and $\hat{\delta}^h_1$ underestimate true motion, whereas $\hat{\delta}^h_2$ overestimates it. Note that motions are not centered around 0 px since they are estimated with respect to the first frame; their temporal mean are practically removed to get centered motions.

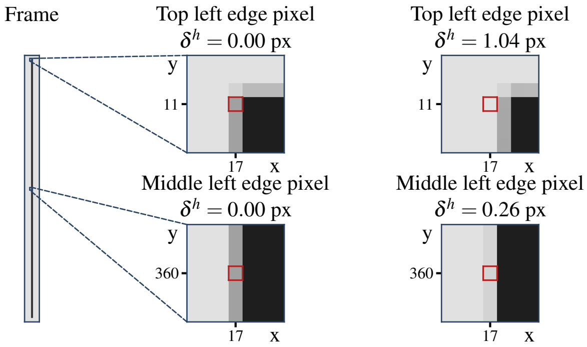

Fig. 2. Edge pixels of interest at equilibrium (left part) and during movement (right part) for motion estimation.

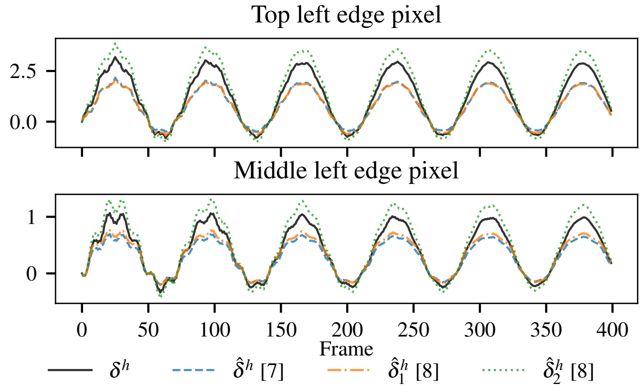

Fig. 3. True and estimated motions (px) for $f = 10.47\,\text{N}$.

3.2. Robustness against motion amplitude

We perform a sensitivity study on motion amplitude to highlight sub-pixel efficiency. To this end, we generate nine videos with an input force f ranging from $0.04\,\text{N}$ to $10.47\,\text{N}$ according to a logarithmic step (see Table 1). We then compute the Pearson correlation coefficient between δ^h and estimated motion $\hat{\delta}^h$ (or similarly $\hat{\delta}^h_1$ or $\hat{\delta}^h_2$) at top and middle edge pixels as:

$$C_{\delta^h \hat{\delta}^h} = \frac{\sum_t \delta^h[t]\hat{\delta}^h[t]}{\sqrt{\sum_t \delta^h[t]^2}\sqrt{\sum_t \hat{\delta}^h[t]^2}}. \qquad (19)$$

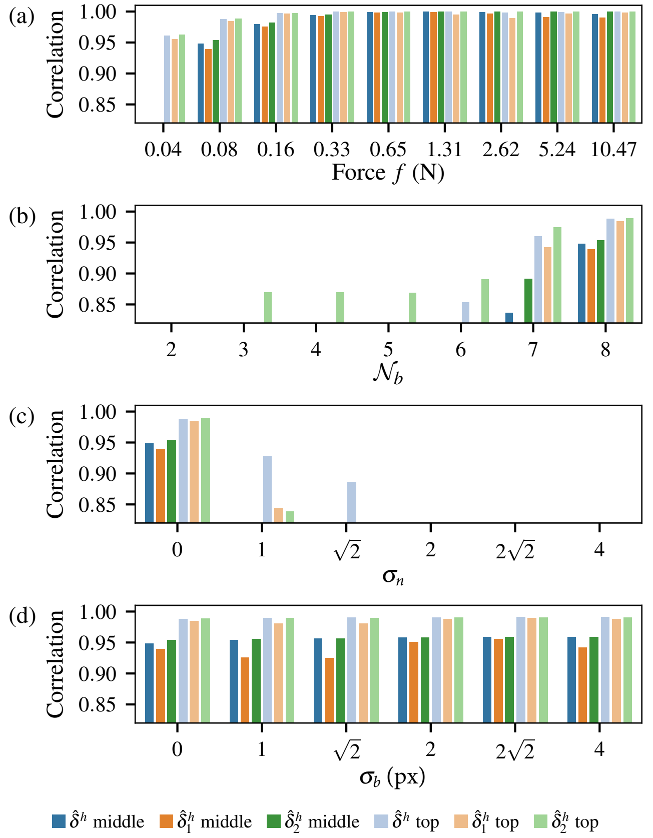

Fig. 4. Correlation between estimated and true motions vs. force (a), or, when $f = 0.08\,\mathrm{N}$, vs. number of quantization bits (b), noise (c) and blur (d) standard deviations.

The results are displayed on Fig. 4(a). Only correlations higher than 0.8 are shown to highlight relevant estimations.

When f ranges from $0.08\,\mathrm{N}$ to $5.24\,\mathrm{N}$, estimators $\hat{\delta}^h$ and $\hat{\delta}_2^h$ reach similar high correlations, whereas $C_{\delta^h\hat{\delta}_1^h}$ is slightly lower. No motion is correctly estimated at middle edge pixel for small motion amplitude ($f = 0.04\,\mathrm{N}$).

3.3. Robustness against gray level quantization

Since outside illumination is not controlled, image contrast may vary. Therefore, we quantize gray levels on $\mathcal{N}_b \in [\![2, 8]\!]$ bits before motion estimation. This study is performed on two videos with true motions of large and small amplitudes generated by input force f of $2.62\,\mathrm{N}$ and $0.08\,\mathrm{N}$.

The three methods estimate large motions with a correlation higher than 0.95 when $\mathcal{N}_b \geq 4$ (no figure for $f = 2.62\,\mathrm{N}$). For small amplitudes ($f = 0.08\,\mathrm{N}$), $C_{\delta^h\hat{\delta}_2^h} \geq 0.9$ when $\mathcal{N}_b = 7$, whereas $\mathcal{N}_b = 8$ bits are necessary to estimate small motions by $\hat{\delta}^h$ and $\hat{\delta}_1^h$ (see Fig. 4(b)).

3.4. Robustness against noise

For this study, Gaussian noise with standard deviation σ_n is added to each video frame ($\mathcal{N}_b = 8$ bits) before motion estimation.

For large amplitudes, correlations are always close to 1.0 for both pixels (no figure for $f = 2.62\,\mathrm{N}$). For small amplitudes ($f = 0.08\,\mathrm{N}$), $\hat{\delta}^h$ is more robust against noise than $\hat{\delta}_1^h$ and $\hat{\delta}_2^h$ (see Fig. 4(c)). This figure shows that the WLS estimator successfully reduces the impact of noise perturbation thanks to the Gaussian spatial constraint.

3.5. Robustness against blur

As the distance between the camera and outside mechanical structure can vary, optical setup may not be always optimal. Therefore, we also study robustness against blur by applying a Gaussian filter (with varying standard deviation σ_b) on video frames.

All methods succeed in estimating large motion with correlations close to 1.0 (no figure for $f = 2.62\,\mathrm{N}$). Figure 4(d) shows that $\hat{\delta}^h$ and $\hat{\delta}_2^h$ provide similar estimations of small motions ($f = 0.08\,\mathrm{N}$) whatever σ_b. At both edge pixels, $C_{\delta^h\hat{\delta}_1^h}$ is the lowest correlation, whatever the blur level.

4. CONCLUSION

In this paper, we compare the performances reached by phase-based motion estimators in terms of correlation with true subpixel motion. For this purpose, synthetic videos are computed thanks to a physical model that simulates vertical cantilever beam vibrations. Beam vibrations cause small or large motion, according to the considered element location along the beam. Since motion amplitude gets larger towards the beam free end, motion is estimated at top and middle edge pixels. Motion is either estimated by a multi-scale approach using a WLS solution to merge phases at different scales or by a fast single-scale horizontal estimator. The latter approach provides the worst estimations on first scale in most cases. On second scale, however, it globally provides the best estimations and is the most robust to gray scale quantization. The multi-scale estimator takes several motion directions into account. It also gives good results and is the most robust against noise while being the most time-consuming. Therefore, this experimental study shows that selection of a relevant scale is the key problem to estimate motion with varying amplitude. However, as this study focuses on horizontal displacement, performances should be assessed when no assumptions are made about motion direction. For the purpose of future vibration analysis of outside mechanical structures, these experiments will be coupled with operational modal analysis methods.

5. REFERENCES

[1] Anders Brandt, *Noise and vibration analysis: signal analysis and experimental procedures*, John Wiley & Sons, Ltd, 2011.

[2] Sung-Wan Kim and Nam-Sik Kim, "Multi-point displacement response measurement of civil infrastructures using digital image processing," *Procedia Engineering*, vol. 14, pp. 195–203, 2011.

[3] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–8, Aug. 2012.

[4] Jaka Javh, Janko Slavič, and Miha Boltežar, "The subpixel resolution of optical-flow-based modal analysis," *Mechanical Systems and Signal Processing*, vol. 88, pp. 89–99, May 2017.

[5] Jau-Yu Chou and Chia-Ming Chang, "Image motion extraction of structures using computer vision techniques: A comparative study," *Sensors*, vol. 21, no. 18, Sept. 2021.

[6] Justin G. Chen, Neal Wadhwa, Young-Jin Cha, Frédo Durand, William T. Freeman, and Oral Buyukozturk, "Modal identification of simple structures with high-speed video using motion magnification," *Journal of Sound and Vibration*, vol. 345, pp. 58–71, June 2015.

[7] Neal Wadhwa, Justin G. Chen, Jonathan B. Sellon, Donglai Wei, Michael Rubinstein, Roozbeh Ghaffari, Dennis M. Freeman, Oral Büyüköztürk, Pai Wang, Sijie Sun, Sung Hoon Kang, Katia Bertoldi, Frédo Durand, and William T. Freeman, "Motion microscopy for visualizing and quantifying small motions," *Proceedings of the National Academy of Sciences*, vol. 114, no. 44, pp. 11639–11644, Oct. 2017.

[8] Yongchao Yang, Charles Dorn, Tyler Mancini, Zachary Talken, Garrett Kenyon, Charles Farrar, and David Mascareñas, "Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification," *Mechanical Systems and Signal Processing*, vol. 85, pp. 567–590, Feb. 2017.

[9] David J. Fleet and Allan D. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, Aug. 1990.

[10] Eero P. Simoncelli and William T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Procs. International Conference on Image Processing*, Washington, DC, USA, oct 1995.