Amplitude measurement of small displacement using video magnification

1st Dominik Ricanek
dept. of Control and Instrumentation
Faculty of Electrical Engineering and Communication, Brno University of Technology, Czechia
ORCID: 0000-0001-5031-2481

***Abstract*—Video magnification algorithms show promising results when used to amplify small vibrations. Measuring these small vibrations is integral in modal analysis of an object and is usually done using specialized vibrometers or accelerometers. Computer vision (CV) systems fall short in this task as the magnitude of the vibrations decreases because of a small SNR (Signal-to-Noise Ratio). In this paper we try to further improve the accuracy of the CV approach by adding video magnification into the image pre-processing stage, allowing the algorithms to measure even imperceptibly small vibrations. For this purpose, we have gather ground truth data with a laser vibrometer in tandem with high speed camera footage of a metal cantilever beam, vibrating in its first mode, and have trained a convolutional neural network for regression.**

***Index Terms*—EVM, phase-based, optical flow, PDV, ResNet18**

I. Introduction

Camera based systems (CBS) remain underutilized in modal analysis of complex objects where transducers and laser vibrometers surpass them in both spatial resolution and frequency range. However, recent advancements in video magnification, specifically the phase-based approach by [12], allow us to improve the spatial resolution of cameras dramatically. This combined with the other benefits of CBS, such as the ease of use and remote sensing, can help them set up a niche as a fast modal shape analysis tool, used to identify places on an object where more precise sensors could be placed.

Using a laser vibrometer, we can measure surface vibrations of the object by scanning. However, this limits us to measurements parallel to the laser's plane and requires expensive equipment. In CBS the entire image plane is used to measure vibrations and, depending on the requirements, the equipment can be affordable. The disadvantage lies in frequency range as a high camera frame rate is naturally going to be more difficult to achieve than high measuring frequency of an accelerometer.

The idea to use cameras to amplify invisible motion dates back to [7] and their Lagrangian magnification. Since then major advancements have been made, starting with the invention of Eulerian magnification by [14] which was then further improved by [12] by using phase differences to allow for greater magnification factors without affecting the SNR. The names of these two algorithms stem from fluid mechanics and allude to their different perspectives; whereas Lagrangian magnification focuses on the motion of individual pixels - as one would particles within a fluid - the Eulerian perspective takes into account the motion of all particles globally through the use of spacial filters.

The method has been experimentally proven to magnify real small motion by [2], where the team found the lowest bound of sub-pixel motion to be 0.00025 pixels. The lower bound can be further improved by using specialized image sensors with greater bit depths or higher contrast.

Attempts have also been made to incorporate phase-based Eulerian video magnification (EVM) into edge devices by implementing it in C++ [3]. These are usually still run on a laptop computer, however. A truly edge computing application is used for example for baby breathing monitor in [5]. Phase-based video amplification nevertheless remains highly computationally demanding [9] and sensitive to larger motion within the picture.

II. Components

Phase-based EVM can be thought of as a combination of three smaller algorithms. That is:

A. Phase-based optical flow

Optical flow is a tool used in computer vision to gain motion vectors of objects within analyzed video. The phase-based EVM does this by looking at phase differences between two frames, rather than intensity differences, as in the seminal algorithm.

The method itself has been around for over two decades but remains sparsely used relative to the other computer vision algorithms [2]. Apart from EVM, it is used for example in autonomous navigation, movement detection and tracking.

B. Image pyramid

An image pyramid decomposition is an over-complete transformation which results in a sequence of progressively downscaled versions of the original image called "levels" as can be seen in Fig. 1. When applied to a video, decomposition is performed on each frame individually. This requirement is the main cause behind the high computational demands of video magnifying algorithms. Indeed one of the major advancements in video magnification was the introduction of Riesz pyramids

The completion of this paper was made possible by the grant No. FEKT-S-23-8451 - "Research on advanced methods and technologies in cybernetics, robotics, artificial intelligence, automation, and measurement" financially supported by the Internal science fund of Brno University of Technology.

as directional filters by [13], which led to a major increase in processing speed.

In our implementation we choose the usual approach of down-scaling by a factor of two between each level and using Gaussian filter to pool the pixels. This way we are left with multiple levels of spatially filtered images which helps us magnify motion beyond just the sharpest edges.

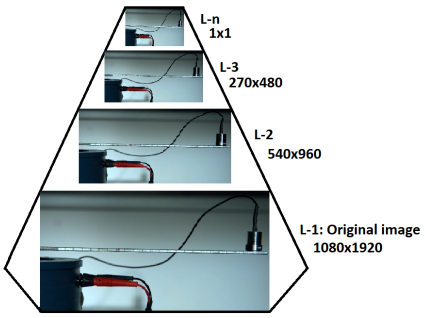

Fig. 1. A general image pyramid with n-levels.

C. Complex-steerable filters

By filters here we mean 2D convolution masks; complex because they are applied in the Fourier domain and thus as complex numbers; and steerable because their response is unaffected by the direction they are applied in - we use eight directional filters derived from a common complex-steerable filter bank as seen in Fig. 2.

In essence these function as edge detectors applied on each transformed frame of the video e.i. each pyramid. For this reason the filters themselves have to be pyramid form. These are sometimes called Riesz pyramids for their special ability to create orthogonal quadrature pairs upon application and thus reducing the number of required directional filters by half - notice the lack of opposing directional filters in our example.

Another useful property of complex-steerable filters are their non-aliased sub-bands thanks to which we can then spatially move the filter response without artefacts. It should be noted that steer-ability is not exploited. [12] [10]

III. Phase based video magnification

One of the consequences of the Fourier theorem is that signals undergoing translation will have some of their components change phase and vice versa - changes in phase of a component sine wave correlate to motion of the signal. This can be easily imagined on a 1D example in Fig. 4 where a wavelet moves in the x axis based on the changes to its phase. This process was shown to have promise in video processing application and the team behind [14] utilized it to avoid the computationally demanding calculation of motion vectors of other video magnifying methods and instead directly manipulate the phase of a video through complex-steerable filters.

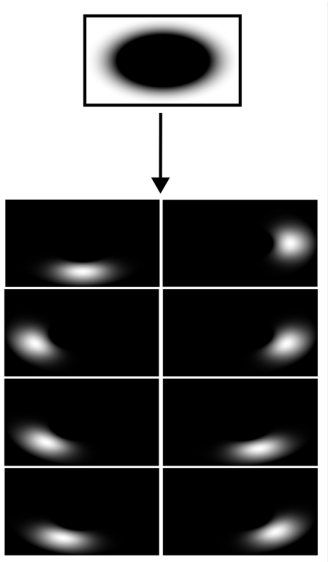

Fig. 2. The complex steerable filter (above) and its bank of eight directional filters.

In practical application we will start by shooting a video with the Nyquist–Shannon theorem in mind. That is at twice, or more, the frame rate of the expected frequency. Note that methods exist to remedy distortion and aliasing caused by capturing at lower frame rates but those were out of scope of this work.

The video is then then split into individual frames and each frame is transformed into an image pyramid using a Gaussian filter. The complex-steerable filters, which will be applied in the next step, are also transformed. However, since the filter bank does not change, these transformations can be done beforehand to reduce processing time. These filters are also already in the Fourier domain in order to utilize the Convolution theorem and further save computational time.

Each frame is transformed into the Fourier domain, centered, and multiplied by each of the eight complex-steerable filters. The result of this process is a new image pyramid composed of complex matrices from which we can calculate the phase as the angle between the real and imaginary parts.

The first frame of the video is chosen as the reference. It is transformed as stated above and used to calculate the difference in phase between all the other frames. These differences are then filtered temporally to band-pass the desired frequencies and multiplied by the amplification constant α (where $\alpha = 1$ corresponds to 100% motion amplification).

The amplified image pyramids are then collapsed and added to the original frames. This way we end up with a synthetic video with amplified motion, without the need to track movements like other video magnification techniques do [12].

IV. Related research

Eulerian Video Magnification has progressed a long way since its debut at MIT in 2012 [14] and nowadays these sorts of algorithms achieve accurate real time amplification of tiny

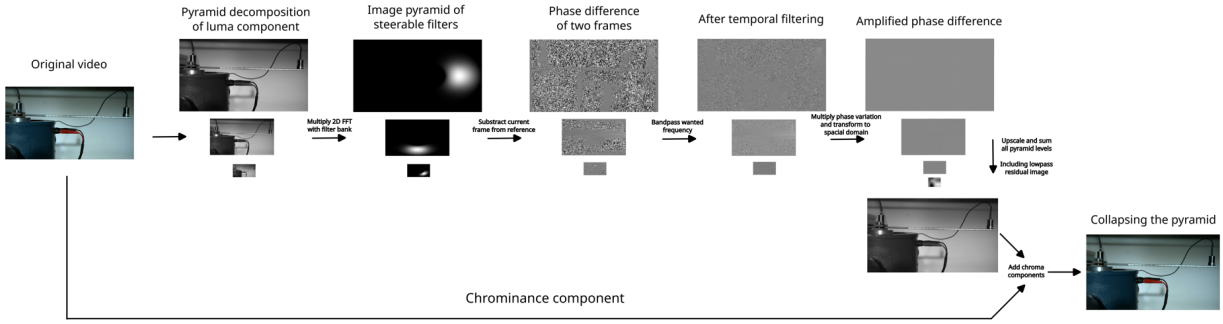

Fig. 3. Full processing chain of the phase video amplification algorithm. Left to right: original image, pyramid decomposition, convolution with complex-steerable filters, phase difference multiplication, temporal filtering, collapsing the pyramid and adding to original frame.

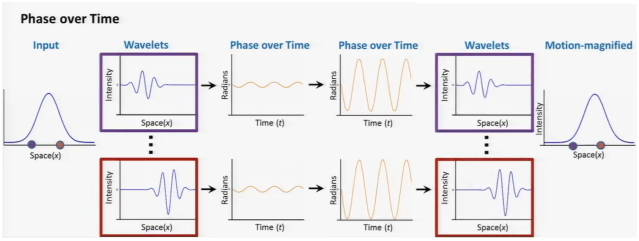

Fig. 4. Translation of a wavelet by changes in phase of its components [11].

motions - be it periodic vibrations or short impulses - in a video. Using spatial phase, it is possible to highly negate the effect of noise within the image on the magnified result [12] and attenuate unwanted movements. This can and has been used to ameliorate the fatal flaw of EVM, large motion [17]. Computer vision systems have been used for precise modal analysis of complex structures such a three storied model of a building [16]. Efforts have also been made to transfer the EVM algorithm into the realm of machine learning, with mixed results [8]. However, to our best knowledge, nobody has yet attempted to use phase-based EVM algorithm to aid a machine learning models in determining the amplitude of the detected vibrations.

When it comes to vibrometry using regression CNNs, the idea has been extensively explored. In [6] it has been used to detect anomalies in rotating machinery and predict bearing faults with an RMSE of 0.97, the authors use 1D convolution of an attached accelerometer in combination with a gated recurrent unit (GRU). Although somewhat unrelated, we mention this here to point out that gathering of raw vibrational data has great potential not only for immediate diagnostic purposes but also for overall quality assessment of the machinery and so a need exists for precise vision based solutions that could replace tactile sensors in the future. A slightly more complex approach has been done by [1] and [18], where 2D CNNs have been used on time-frequency data images for surface roughness estimation and fault diagnosis. In the case of [1] a classifier is then trained to also differentiate between various types of bearing faults or detect tool wear on a milling machine bit. The authors of [18] use continuous wavelet transform (CWT) instead of short-time Fourier transform (STFT) and a nonlinear auto-regressive neural network to generate more signals to tackle the problem of imbalanced datasets. In [4] vibration data is combined with sound samples to allow for fault detection as well as monitoring of the machine's state using fast Fourier transform and principal component analysis (PCA) to reduce the amount of data. This shows that a camera could be used as a two in one type of sensor and opens the avenue for even more opportunities. In some cases, old archived footage from plant cameras could be used to train a more robust model. In [15] the authors use a high speed industrial camera in a well controlled setting to capture, detect and automatically analyze a rotating body. They use a re-identification network which proved to remove the jitter otherwise associated with frame-to-frame automatic, anchor-free region of interest locating.

V. Experimental setup

The dataset in this paper has been gathered as a part of a larger experiment. We use three videos, with increasing amplitude of vibration, of a double sided cantilever beam actuated by a vibrating shaker connected to a function generator Siglent SGD 2042X through an amplifier B&K Type 2732.

The right side of the beam is 30 cm long and has a resonant frequency of 33.8 Hz along the first mode. This frequency was chosen because the part of the broader experiment from which we source our data was intentionally limited to capturing at 120 frames per second. That being said, we used a high speed camera Phantom v1840, again limited to 120 frames per second in this particular case, because we needed a wider range of frame rates to be available. We used Canon EF 20-70mm f/2.8 II USM lens and the entire setup was illuminated by high powered LEDs controlled by the GSVitec MultiLEDG8 driver as can be seen in Fig. 5.

The ground truth measurements at each of the annotated points in Fig. 6 were taken with a laser vibrometer Polytec

PDV 100 and data logged using LabVIEW. Note that the sensor at the end of the arm in Fig. 6 is a charge accelerometer and is not relevant to this study.

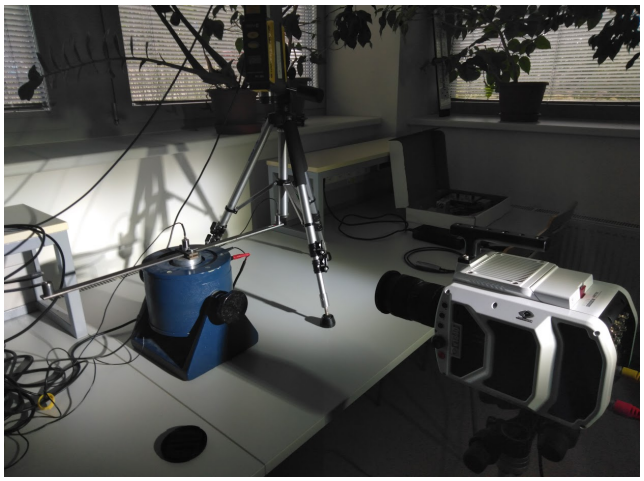

Fig. 5. Experimental setup. Left to right: shaker with cantilever, laser vibrometer, Phantom v1840. This setup shows measurement of the ground truth for one of the 28 points.

Each point along the arm had first been measured by the laser accelerometer before the vibrating arm had been shot at the aforementioned 120 frames per second by the Phantom V1840 camera in raw format. This way we collected three 5 s clips. The distance of the arm from the camera lens was always 50 cm and the exposure was set to 100 μs.

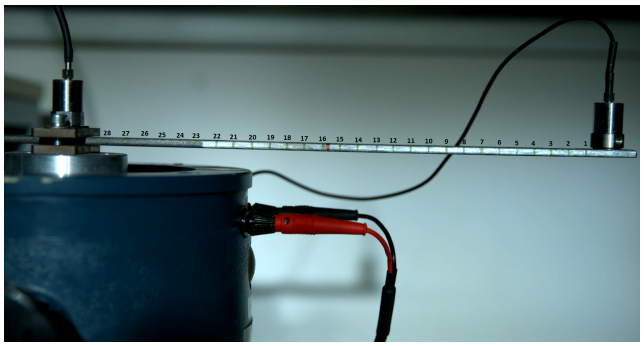

Fig. 6. Labeled arm points that were measured by the laser vibrometer. Spaced apart by 1 cm.

VI. Image pre-processing

As we are dealing with amplitude measurements of a temporal signal, the video must first be transformed into a temporal image. In our case we take a 1 pixel wide vertical slice from each frame around the point to be measured, and stack them side by side to create a time-slice image, as can be seen in Fig. 7 on the right.

Before we settled on using a regression CNN, we tried two different classical CV approaches. All our experiments were compared to ground truth data measured with a laser vibrometer.

At first, since we mainly focus on observing the edge of the cantilever, we tried implementing an edge detector to get an amplitude envelope from the time-slice image. However, since the amplitude envelope of our time-slices does not have clear enough edges, tuning an edge detector proved to be too complicated.

The second attempt included applying a 2D Fourier transform on the time-slice images and looking for peaks in the frequency domain. This approach had good results but required manual selection of the frequency peaks which was not possible to automate.

Thus we settled on using a regression CNN. The data further proved to be easy to augment by doing linear interpolation between each pair of the 28 measured points. We gained around 1400 image-amplitude pairs (the distance in Fig. 6 from point 28 to point 1 spans 1400 pixels) from each video, resulting in a dataset containing 4212 samples.

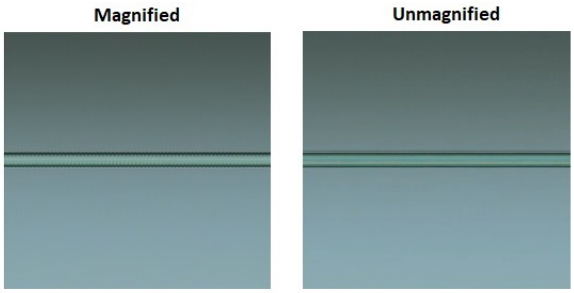

Fig. 7. Comparison of non-magnified and magnified point number 2 from the dataset with the lowest amplitude

VII. CNN training

We tested several different architectures. First we tried a naive approach and created a simple CNN with two convolutional layers and three fully connected layers, which took in 32x32 images of the cropped time-slice. However, the network was not deep enough to provide good feature selection and ultimately gave random results.

Then we tried training an older deep CNN architecture from scratch. For this purpose we chose VGG16 which accepts 244x244 images, making it potentially usable for lower frequency vibrations. But training the architecture from scratch provided results with very high error.

We retained the 244x244 input size and decided to leverage transfer learning instead of randomizing initial weights. At the same time we decided to try two other architecture, AlexNet and ResNet18, from which we chose the latter as it showed best performance during training.

Trying the more complex versions of ResNet provided only marginal improvements, so we stayed on ResNet18. We tested training with three different final layers of increasing depths, however, the simplest one gave the best results; it consists of a single linear layer, reducing the input 512 features into a single output, and a sigmoid function. We can use the sigmoid

function because out data fits into the range of 0 to 1, however, if higher displacement is present in the training dataset, ReLU could be used instead. Finally, we used mean squared error as our loss function to train for regression.

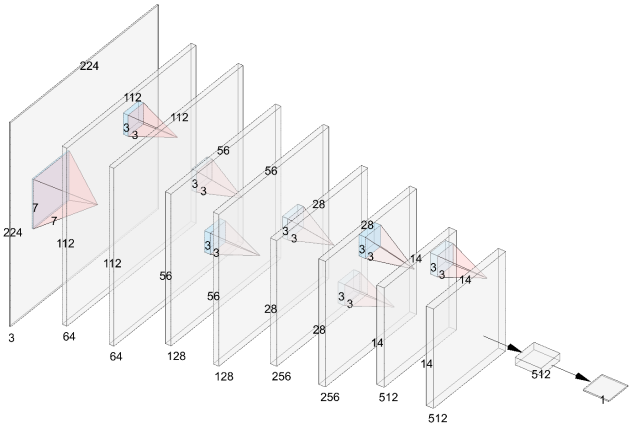

Fig. 8. CNN architecture, ResNet18 adjusted for regression, final layer output is passed through the sigmoid function.

VIII. Results

The CNN performed well without EVM pre-processing for two of our three videos where SNR was low enough for the signal to get picked up by the network. However in the last dataset the noise was too much for our network to handle, as can be seen in Fig. 9.

In Fig. 10 we used phase-based EVM in pre-processing before creating time-slices and passing them to the network. The amplification factor was chosen heuristically to be $\alpha = 100$ and the network outputs were divided by the same number to get the final measurement.

While the result of our CNN and the ground truth of our laser measurements generally correlate, there are still major deviations from the true values. However, the result while using EVM is much better than without using it.

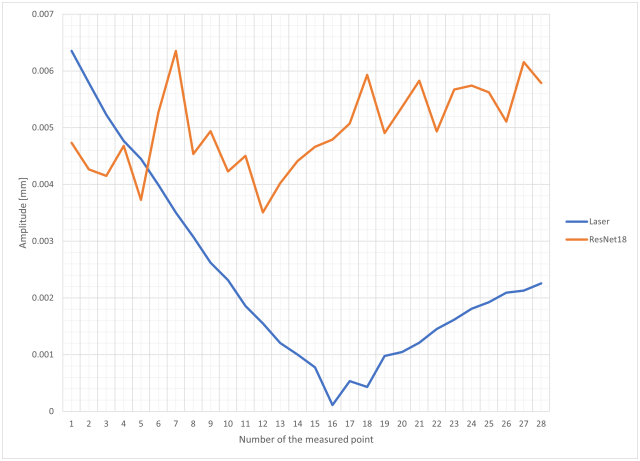

Fig. 9. Measurement of amplitude from video without EVM pre-processing

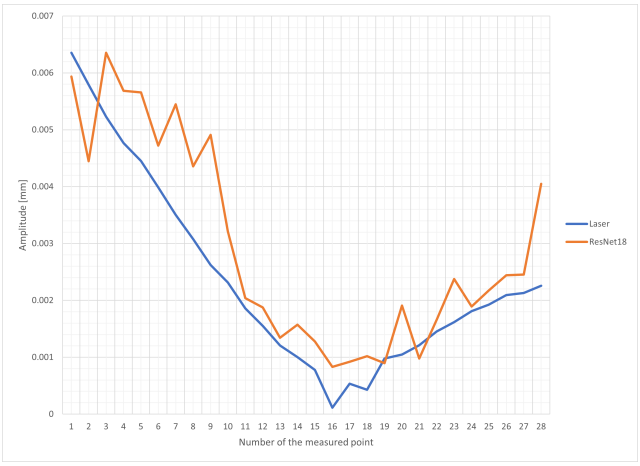

Fig. 10. Measurement of amplitude from video with the aid of EVM

Point	Laser [μm]	ResNet18 [μm]	Δ [μm]
1	6.353	5.933	0.420
2	5.790	4.445	1.345
3	5.231	6.354	1.122
4	4.765	5.686	0.920
5	4.448	5.656	1.207
6	3.989	4.723	0.733
7	3.501	5.448	1.946
8	3.076	4.356	1.279
9	2.621	4.911	2.289
10	2.314	3.211	0.897
11	1.859	2.041	0.182
12	1.551	1.876	0.324
13	1.205	1.342	0.136
14	1.000	1.571	0.571
15	0.775	1.276	0.501
16	0.111	0.831	0.719
17	0.535	0.921	0.386
18	0.429	1.016	0.586
19	0.977	0.897	0.080
20	1.046	1.906	0.860
21	1.211	0.976	0.235
22	1.454	1.658	0.204
23	1.617	2.376	0.758
24	1.809	1.893	0.084
25	1.925	2.176	0.250
26	2.092	2.442	0.349
27	2.129	2.453	0.323
28	2.255	4.050	1.794

TABLE I
COMPARISON BETWEEN LASER MEASUREMENT AND CNN REGRESSION FOR LOWEST AMPLITUDE DATASET

IX. Conclusion

The goal of this paper was to familiarize the reader with the possibilities of phase-based Eulerian video magnification when applied as an auxiliary algorithm to other computer vision algorithms.

We explained the three major parts of phase-based EVM and how they work together to apply linear magnification of the selected motion within a video.

Furthermore, we trained a ResNet18 as a regression CNN, utilizing transfer learning to measure vibration amplitude of a point from a time-slice image of the video.

We used three videos of a resonating cantilever beam to create a dataset containing 4212 images.

The resulting CNN worked well on two of the three videos but struggled with the third, where vibration amplitude was overwhelmed by noise. However, after applying EVM in the pre-processing step, the network's outputs correlated relatively well with the ground truth data.

In future work we will focus on creating a CV algorithm to extract features of the vibration signal from the video and pass them to a more specialized neural network to achieve better results than the more general oriented CNNs can offer.

REFERENCES

[1] Han-Yun Chen and Ching-Hung Lee. Deep learning approach for vibration signals applications. *Sensors*, 21(11):3929, 2021.

[2] Sean Collier and Tyler Dare. Accuracy of phase-based optical flow for vibration extraction. *Journal of Sound and Vibration*, 535:117112, 2022.

[3] Nikolaos Giannopoulos. Phasebasedevmcpp. https://github.com/NikolaosGian/PhaseBasedEVMCpp, 2023.

[4] Seulki Han, Nasir Mannan, Daryl C Stein, Krishna R Pattipati, and George M Bollas. Classification and regression models of audio and vibration signals for machine state monitoring in precision machining systems. *Journal of Manufacturing Systems*, 61:45–53, 2021.

[5] Luke Hsiao. Cribsense. https://github.com/lukehsiao/CribSense, 2016.

[6] Kwangsuk Lee, Jae-Kyeong Kim, Jaehyong Kim, Kyeon Hur, and Hagbae Kim. Cnn and gru combination scheme for bearing anomaly detection in rotating machinery health monitoring. In *2018 1st IEEE International conference on knowledge innovation and invention (ICKII)*, pages 102–105. IEEE, 2018.

[7] Ce Liu, Antonio Torralba, William Freeman, Frédo Durand, and Edward Adelson. Motion magnification. *ACM Trans. Graph.*, 24:519–526, 07 2005.

[8] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018.

[9] Karl Pauwels and Marc M. Van Hulle. Realtime phase-based optical flow on the gpu. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[10] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

[11] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Trans. Graph.*, 32(4), jul 2013.

[12] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32, 07 2013.

[13] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Riesz pyramids for fast phase-based video magnification. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2014.

[14] 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 - TOG*, 31, 07 2012.

[15] Rongliang Yang, Sen Wang, Xing Wu, Tao Liu, and Xiaoqin Liu. Using lightweight convolutional neural network to track vibration displacement in rotating body video. *Mechanical Systems and Signal Processing*, 177:109137, 2022.

[16] 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*, 85:567–590, 2017.

[17] Yichao Zhang, Silvia L. Pintea, and Jan C. van Gemert. Video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] Quan Zhou, Yibing Li, Yu Tian, and Li Jiang. A novel method based on nonlinear auto-regression neural network and convolutional neural network for imbalanced fault diagnosis of rotating machinery. *Measurement*, 161:107880, 2020.