Title

Fourier Representations with Sequentially-Trained, Shallow Neural Networks for Real-Time Odometry and Object Tracking

Author

Rodriguez, Frank Christopher

Publication Date

2022

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Fourier Representations with Sequentially-Trained, Shallow Neural Networks for Real-Time Odometry and Object Tracking

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Mechanical Engineering

by

Frank C Rodriguez

September 2022

Dissertation Committee:

Dr. Luat T. Vuong, Chairperson
Dr. Jun Sheng
Dr. Jonathan Realmuto

The Dissertation of Frank C Rodriguez is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Thank you to my advisor, Dr. Vuong, for always believing in me and providing invaluable guidance, support and encouragement over the course of this project. Thank you to Baurzhan Muminov, Altai Perry and Xiaojing Weng for your mentorship and support in the lab. Thank you to the students who have been a part of the Vuong Lab for your insightful conversations and help.

I would also like to thank my wife Gabby for all her support throughout the course of my studies and my family and friends who have stood by me through it all.

ABSTRACT OF THE THESIS

Fourier Representations with Sequentially-Trained, Shallow Neural Networks for Real-Time Odometry and Object Tracking

by

Frank C Rodriguez

Master of Science, Graduate Program in Mechanical Engineering
University of California, Riverside, September 2022
Dr. Luat T. Vuong, Chairperson

Fourier-domain correlation approaches have been successful in a variety of image comparison approaches. However, these correlation approaches also lose performance when patterns, objects or scenes in images are distorted. Current Fourier correlation approaches also require high-power in order to produce accurate results. With our approach we utilize Fourier-domain preprocessing with shallow neural networks to infer the 3-D movement or position of the camera relative to an object or scene. This approach enables us to demonstrate potential for novel Fourier-plane cameras, which use sequential frames for visual odometry. We also propose a potential future study for a hybrid vision "event camera" system capable of position inference by using an optical encoder imaged in the Fourier-plane.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Introduced in 1975, Fourier-domain image processing approach, known as phase correlation (PC), has been used for a range of applications. The applications of PC include image analysis, object detection and classification, pattern comparison, visual odometry (VO) and motion estimation [1, 2, 3, 4]. PC approaches have advantages over other correlation techniques, which include improved resolution and computational efficiency, insensitivity to brightness, contrast and frequency-limited corruptions and results independent to scene and objects [5, 6, 7, 8, 9].

Theoretically, Fourier-domain PCs provide information about the scaling and rotations of image patterns[3, 10, 11]. However, a major drawback of Fourier-domain methods is their sensitivity and susceptibility to frequency-dependent distortions, which are caused by disjoint segments between the two images and vibration or jitter. Approaches to address this issue include space-domain windowing [12], frequency filters [10, 13], sampling, and image gradients [14]. This latter method enables robust estimations of the translation, rotation, and scaling appearing as spatial displacements between pairs of images.

Currently, motion estimation approaches are more commonly accomplished via deep learning, which has rapidly been adopted for VO applications such as scene tracking or optical flow for unmanned aerial vehicle (UAV) and robotics navigation [15, 16, 17], many of which incorporate Fourier-domain or PC methods [18, 19, 20, 21, 22]. Supervised [23] and unsupervised [24] learning approaches have been demonstrated to predict the speed, depth, and position of system objects, even with a monocular imaging system. Although current motion estimation techniques can be applied accurately, a major limiting factor is the speed of the algorithms during implementation [25]. The state-of-the-art techniques achieve 40-60 frames per second (FPS) [26, 27], however this raises questions of practicality, especially in real-time applications for obstacle avoidance algorithms, which aim to use this sensor data for further inference [28].

Another concern with the current state-of-the-art approaches is the high power consumption that is required for VO algorithms implementing deep learning, particularly for remote-powered applications. In [29], the 2D motion problem is estimated by 1D Markov chains to simplify and parallelize analysis before the neural network, however this method requires a 250-W GPU. Wang *et al.* implement pyramidal processing, spatial warping, and a cost volume estimate for high-accuracy, large spatial disparity optical flow calculations, however this approach uses 8 650-W GPUs. [30]. For applications that involve small-platform, communication-deprived, autonomous, real-time systems, we are interested in VO methods that are able to infer position with low-power. The task-specific engineering of such high-speed, low-power systems would integrate hybrid optical-electronic architectures, to include encoding optics, algorithms and camera sensors or hardware.

Here, we combine Fourier image processing approaches with a shallow neural network. Our efforts run parallel to recent approaches with focal-plane sensor processing units where image preprocessing is integrated into the detection hardware [31]; such units have achieved VO at a rate of 300 FPS [32] using a small laptop CPU. There has also been significant advances with 3-D imaging with optically-encoded imaging or deep optics [33, 34] and hybrid optical-electronic systems [35, 36, 37], which build off extensive research knowledge developed for optical computers from the 1970s [38, 39, 40, 41, 42, 43, 44]. In [33], a siphon-shaped U-net dense neural network [45] is used and in [34], a U-Net with residual learning integrating additional spectral representations is implemented. Both [33] and [34] achieve reconstructed depth maps for lidar-like image reconstruction using chromatic dispersion and color sensors. In contrast, with our approach towards Fourier optical preprocessing with even simpler machine learning [36], we take a step back and study object tracking with a monochromatic source and sensor.

In our work, we build on spectral and Fourier representations of data, which influence convergence in machine learning neural networks [46]. To build axioms for design, we aim to answer, what are the simplest optical encoders needed to deconvolve information from Fourier-plane data? If there is information encoded in the Fourier-plane, where is it most likely to be located and how should it be sampled? With our approach, we avoid deep learning, which represents a strong departure from current efforts with VO. We instead implement computationally simpler models that converge with smaller data sets; this classical approach is like LASSO [47] and regression based algorithms[48], with shallow and dense neural networks. In pairing Fourier preprocessing with machine learning, we find that the simplest neural networks are capable of learning from fringe-like patterns produced by taking the CPS computationally. The approach is successful because simple neural networks are capable of taking the inverse Fourier transform of the CPS to rapidly infer the fringe spacing or shifted movement of spectral features.

The work in this paper is motivated by recent bio-inspired ”event cameras”, which employ a sensor that incorporates temporal data streams to determine the difference between images [49]. Here we describe an approach using the cross power spectra (CPS) of high-speed, high-resolution drone flight simulation video captured by a forward facing camera as shown in Fig. 1.2. The CPS is the normalized Fourier transform of the cross-correlation of two images. With the CPS inputs, we train a regression-based neural network, as opposed to using deep learning, to infer velocity data. We retrain the model with different trajectories and different scenes to achieve generalized predictions for unseen trajectories. While our model achieves high-frame-rate video processing, the image *preprocessing* at 95 FPS is significantly slower than the model inference, which can reach speeds upwards of 52,000 FPS. The computational Fourier-domain pre-processor is able to operate in real time for image capture within 95 Hz, but becomes the bottleneck at 50 orders of magnitude lower than the inference rates of the NN. The error for the sequentially trained Neural Network is 3% and 5% for the no-hidden layer and one-hidden layer model, respectively.

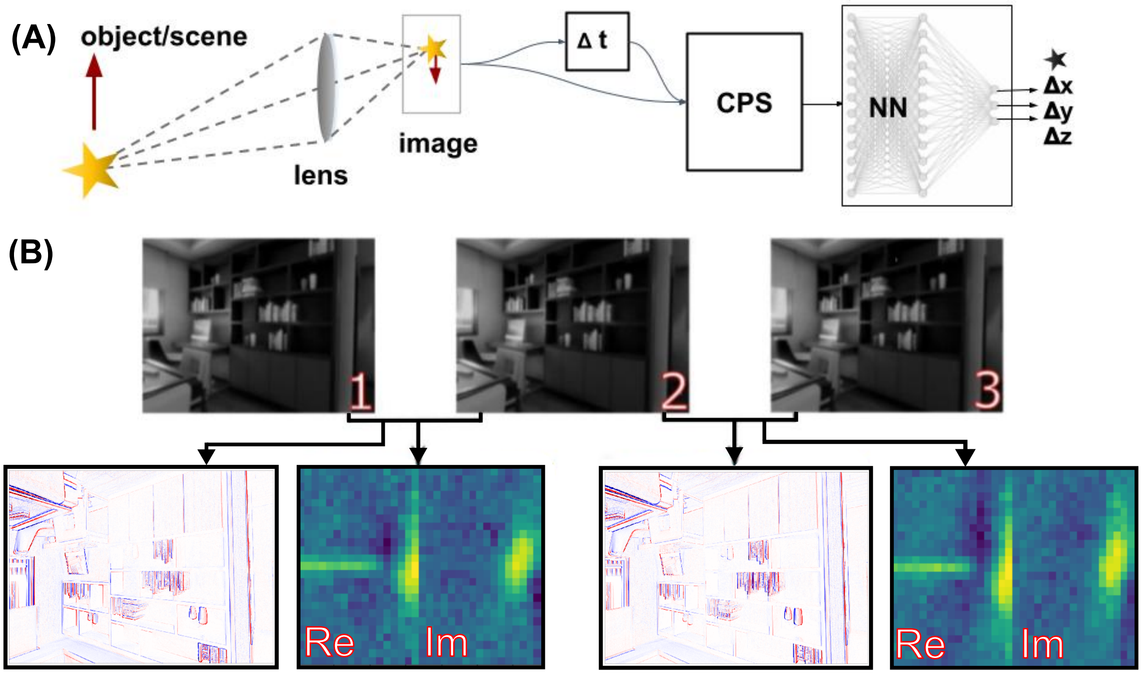

Figure 1.1: A) Schematic showing the pipeline for Experiment 1. The inputs to a shallow, dense, neural network are the cross power spectra (CPS) Eq. 2.1 or the normalized Fourier transform of the cross correlation from sequential images from drone video. B) Illustration of inputs to the simple neural network composed of CPS real and imaginary half-planes. The images here demonstrate sequential frames mid flight, with a change in angle of ¡ 0.06 degrees, which is visible in the downsized CPS. The difference between each pair of sequential frames, as well as the CPS are shown here.

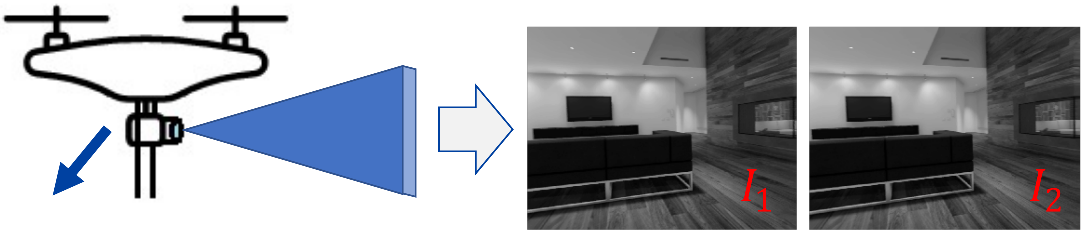

Figure 1.2: Schematic showing potential application of the approach with a drone. Drone captures images, Fourier pre-processor is applied computationally to pairs of images, and a small brain neural network infers position as illustrated in Fig. 1.1.

Chapter 2

Towards a Fourier-Plane Event Camera

2.0.1 General Approach

The Blackbird dataset is an indoor UAV simulation dataset containing several large flight sequences taken at high speeds (120 FPS) [50]. The dataset consists of a total of 18 trajectories across five different environments and includes optical sensors applicable to monocular, stereo and down-facing VO cameras with approximately 60-degree field-of-view, high-definition images [50]. We implement the approach described in Fig. 1.1(A). The difference in position from millimeter-accurate (360 FPS) motion capture data is used to train the neural network. The CPS is generated from sequential frames taken from the Blackbird dataset to estimate 3D motion of the drone. The cross power spectra (CPS) for two 2D images I_1 and I_2 is defined:

$$\mathrm{CPS}(k_x, k_y) \quad = \quad \frac{\mathcal{F}\{I_1(x,y)\}\mathcal{F}^*\{I_2(x,y)\}}{|\mathcal{F}\{I_1(x,y)\}\mathcal{F}^*\{I_2(x,y)\}|}, \tag{2.1}$$

and $\mathcal{F}^*$ is the complex conjugate of the Fourier transform $\mathcal{F}$. The numerator of Eq. 2.1 is the Fourier transform of the 2D cross-correlation between two real images, I_1 and I_2. Given the camera field of view and high-definition images, the CPS is capable of capturing small-tilt changes due to the jitter of the drone before take-off [Fig. 1.1]. The input consists of half of the real component and half of the imaginary component of the CPS stacked side by side and is shown in Fig. 1.1B. Since the input images are all real valued, the resulting CPS is symmetrical and only half of the CPS array is needed to retain the relevant features of the changing scene. Although an edge is produced in the input due to stacking the CPS components, the 2D data is reshaped into a 1D array when input to the neural network. This is one advantage of using spectral representations and shallow neural networks: the order of the input data does not matter. In fact, the shallow neural network does not assign a heavy weight to edges in the in input, but instead the majority of the higher neural network weights are associated with the lower-frequency Fourier components.

The PC method for VO—to determine the offset or spatial disparity between two images—typically refers to the inverse discrete Fourier transform of the CPS, which obtains a Dirac delta function, which shifts in location with the spatial disparity. Conventionally, if there is a simple spatial disparity between the images or features, then

$$I_2(x,y) = I_1(x - x_0, y - y_0), \tag{2.2}$$

and in this specific case, we apply the shift property of the Fourier transform, $\mathcal{F}\{I(x-x_0, y-y_0)\} = \mathcal{F}\{I(x,y)\}e^{-ik_x x_0 - ik_y y_0} = \tilde{I}(k_x, k_y)e^{-ik_x x_0 - ik_y y_0}$, where $\tilde{I}$ is the Fourier transform of the shifted image, in order to achieve a highly compressed signal. The PC classifies the shift with a *scaled* δ-function,

$$\begin{aligned} \text{PC} &= \mathcal{F}^{-1}\{\text{CPS}(k_x, k_y)\} \qquad (2.3) \\ &= A\delta(x + x_0, y + y_0) + \mathcal{F}^{-1}\{\text{CPS}_\Delta\}. \qquad (2.4) \end{aligned}$$

Even with smooth 2D motion and a downward-facing camera, there is power outside of the δ-function magnitude, A, which is likened to noise; for an $N \times N$ CPS, the signal-to-noise ratio $SNR = \frac{NA}{\sqrt{1-A^2}}$ [1]. This noise arises because each image has a different outer boundary, so there are disjoint portions of each image, or portions associated with each image that have no relation to the other, denoted by CPS_Δ. The challenge with applying PC approaches to VO with aerial drones lies largely in the fact that the drone does not strictly translate a uniform distance x_0, y_0; the drone also tilts and moves forward, leading to significant CPS_Δ as well as distortions in the the $\delta-$function. Additionally, rather than using a downward facing camera, it makes sense to have a forward-facing camera that would avoid objects or obstacles that lie in the direction of motion of the UAV. Therefore, we focus on applying Fourier-domain PC methods using forward-facing camera images to infer the 3D motion as shown in Figure 1.2.

2.0.2 Video Preprocessing and Dataset Alignment

Simulation video files and gyroscope data for three flight paths from the Blackbird dataset are used in our approach. Each frame from the video is downsized and appended to an array to ensure the images remain in the correct order over the course of the flight path. Two of the flight paths used traverse through two environments each, giving a total of 5 datasets that are used to sequentially train the neural network model.

The image preprocessing for the neural network is shown in Fig. 2.1 and consists of first downsizing by 50% and applying a Gaussian blur with a kernel size of (21,21) to the video frames. The Fourier transform of each image is taken and convolved with a super-Gaussian filter in order to eliminate high frequency noise. The CPS is then calculated using the filtered Fourier transform of the preprocessed images, and subsequently cropped to a size of 34 by 25, due to the CPS compressing the necessary data into a small section in the low frequency region. The left half of both the real and imaginary component of the CPS are appended into a single matrix to retain the necessary information for inferring the change between each set of frames. The gyroscope data is processed by taking the difference in position between the corresponding frames and normalized on the interval $[0, 1]$.

The gyroscope data from the file is assigned to a dataframe using the Pandas library and the index value is set to the timestamp column to match measurements with the corresponding images. Since the downloaded lengths of the gyroscope data and the timestamp files for each flight path are not synchronized, it is necessary to determine the offset between the gyroscope and video data. Once the offset is determine, six position and orientation values are extracted and the calculated and normalized x, y and z-direction data is fed into the neural network as training data.

The dense neural network is built using a sequential model with a linear activation and mean squared error loss function. The model is also developed by varying preprocessing parameters and comparing the results to determine the highest performing model. Multiple downsizing and cropping parameters were selected and tested in training the model. The

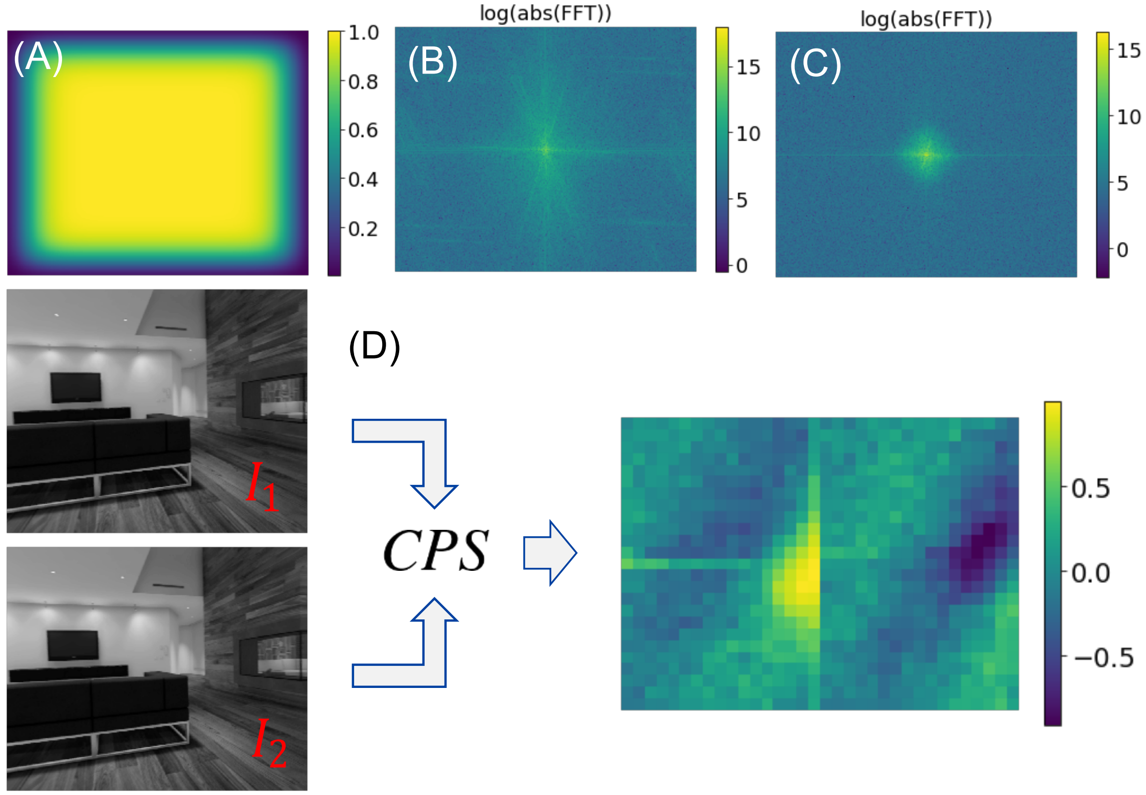

Figure 2.1: (A) The super-Gaussian low-pass filter. (B) Log of the absolute value of the Fourier transform of I1 (C) Log of the absolute value of the Fourier transform of I1 when convolved with the super-Gaussian filter. (D) The pipeline for the preprocessing and computation of the input to the neural network. I1 and I2 are downsized, blurred and the Fourier transform of each image is computed. Each Fourier transform is convolved with the super-Gaussian filter and the CPS is calculated from these filtered Fourier transforms. The left half of each resulting real and imaginary component are cropped and stacked horizontally as the neural network input.

model is trained through each flight path using architectures with both no hidden layers and one hidden layer, the latter of which consisted of the number nodes equal to the product of the input dimensions. The number of training epochs for each model is also kept proportional to the downsizing parameter in pre-processing for each trial by setting the value as the product of the input CPS dimensions. For the CPS generated from video frames from the Blackbird dataset, the model is trained on three unique flight paths set in

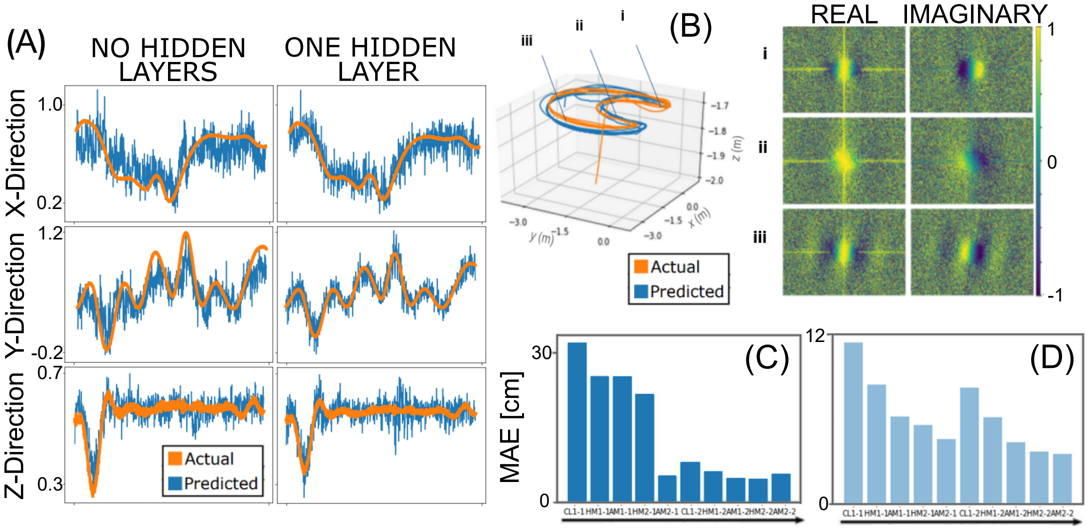

Figure 2.2: A) Prediction results for testing segment of Ampersand flight path, using no hidden layers and one hidden layer model. B) The actual and predicted trajectory of the half moon flight path with the real and imaginary components of the at 3 positions. Convergence of the error after sequential training with different flight paths for C) no hidden layers and D) one hidden layer.

five unique environments. Each model is trained and saved on three flight paths using a validation split of 0.1, which uses 90% of the data to train and 10% to test the model. For the prediction phase, we use only the CPS on a fourth dataset, the model is never ”shown” this dataset.

The neural network model is initially trained with 90% of the Ampersand flight path in a single enfironment. The data is shuffled for training the neural network and the x and y- velocity values are predicted during the testing segment of the neural network. (See Supporting Information for details on the neural network training.) Figure 2.2(A) shows the initial prediction results for the Ampersand flight path. Performance for the z-direction inference improves most significantly of the predicted directions by the inclusion of the single neural-network hidden layer in the model architecture. After a single training on the Half Moon flight path in a single environment, the x and y-direction prediction is high. With one hidden layer, the average error is 9-15% and with no hidden layers, the the average error is 32-54%.

We train the neural network model sequentially using the input CPS from different flightpaths, building upon the previously saved model during each iteration. The error decreases and then flattens, which indicates a coarse generalization. As the model is sequentially trained, it shows better accuracy in the predicted results and a greater ability to reconstruct the flightpath after each iteration. The generalization of the model is shown in 2.2 (C) and (D), which demonstrates the decrease in error for the predicted velocity values over the course of the sequential training. The model with one hidden layer initially outperforms the model with no hidden layers but when retraining on other sets, the error increases again. The model with no hidden layers converges to an error of 5% and the model with one hidden layer converges to an error of 3%.

The 3D flight path reconstruction is shown in Fig. 2.2(B) and is calculated by integrating the predicted velocity and multiplying the value by the change in time between each point and multiplying by a factor of 10^{-6}. This 3D reconstruction demonstrates the effectiveness of our approach in determining the position of the drone using the CPS of the sequential video frames as input to the neural network. As the model is subsequently trained, the error decreases with the Mean Absolute Error (MAE) for the no-hidden and one-hidden layer models shown in Fig. 2.2 (C) and (D) respectively.

Our results indicate that our simple, Fourier-domain preprocessing approach provides a model with generalizability within a few hours on a laptop CPU. One disadvantage

Table 2.1: Summary of results.

	Preprocessing time (ms)	NN Latency (ms)	Hidden Nodes	MAE
No Hidden layers	1	0.025	-	0.06
1 Hidden layer	1	0.06	850	0.04

to our approach is that it is a supervised approach and trains with labeled gyroscope data. Additionally, while our model achieves high-frame-rate video processing, the image *preprocessing* (95 FPS) is up to 500 times slower than the model inference (52,000 FPS), the Fourier-domain preprocessing is the bottleneck in our approach. Table 2.1 shows the reconstruction speeds for our approach.

The key to successful training of our shallow neural network lies in adequate compression and downsizing of images to reduce the signal to noise ratio (SNR) and the minimization of the size of the neural network inputs. With further downsizing, the model is trained faster at the expense of accuracy. (See Supporting Information for details on downsizing and sampling.) Notably, the compressed features reside at locations where the light intensities are very low. In other words, the correlations specific to VO are not present in the direct line-of-sight intensity patterns. With this understanding, we may better sample or ”drop out” sensor data to minimize the back-end electronic processing. Our results advance understanding of object tracking with Fourier-domain PC and CPS methods, and, because they are centered on the axiomatic design, may be useful to the future development of focal plane sensor preprocessors [31].

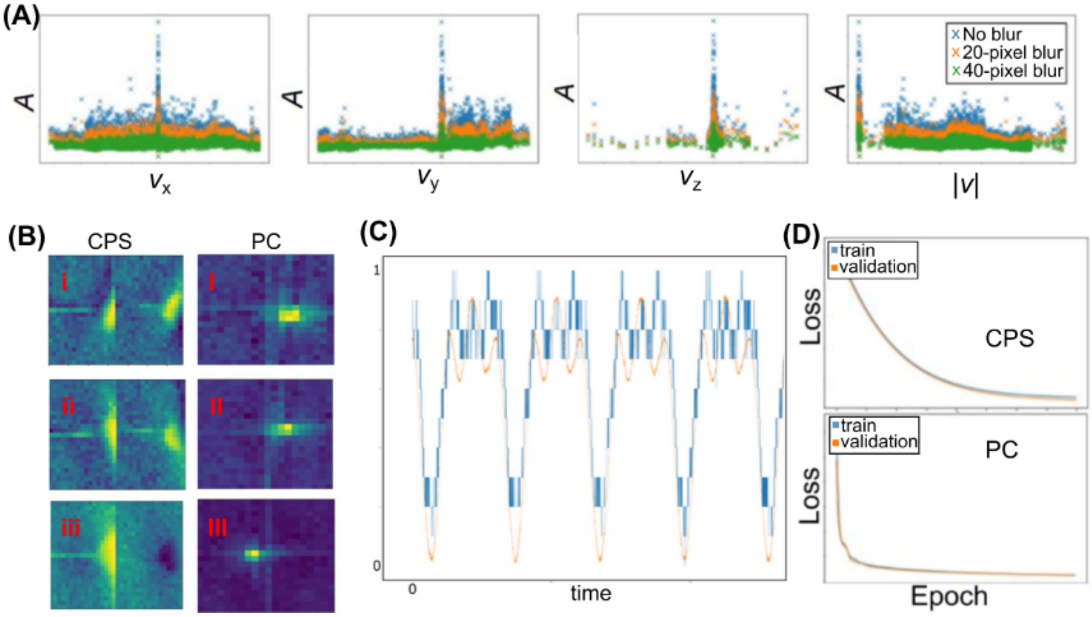

Figure 2.3: Phase correlation (PC) [Eq, 2.4] δ-function peak height, A, with 0, 20, and 40-pixel blur as a function of (A) x, y, and z components of the velocity and the velocity magnitude. (B) Real and imaginary half-planes of the Cross Power Spectra (CPS) and corresponding PC at 3 sample locations (C) Actual and predicted normalized x-velocity values using the maximum pixel position of the PC. (D) Comparison of convergence rates for CPS and PC inputs to the neural network with no hidden layers. The PC converges faster and with lower error due to the higher degree compression of the PC.

2.1 Analysis and PC vs CPS

As previously addressed, the PC is the discrete Fourier transform of the complex-valued CPS, includes a δ-function with amplitude A that representing the shift between the two images [Eq. 2.4]. Figure 2.3A relates the PC peak value amplitude, A and velocity for the Half-Moon flight path images (first environment). The highest values of A occur with low velocities, but this relationship between A and velocity varies with scene and trajectory. In this way, prior analyses with SNR and phase correlation methods are not relevant, since in those prior efforts noise is considered to be independent of the signal [1]. Figure 2.3A tells us that the δ-function peak is heavily distorted with lower amplitude over a larger area when the drone is moving. Three samples of the input are shown in Fig. 2.3 B for the real and imaginary half-planes of the CPS and corresponding PC. Again, the PC is plot for the left half planes of the real and imaginary components of the PC.

Figure 2.3C illustrates this distortion; here we plot the disparity prediction based solely on the location of the highest-amplitude pixel in the PC. The discretization of the prediction reflects the sampling in the PC. What Fig. 2.3C illustrates is that the neural network will not accurately predict position solely from the maximum-value pixel of the PC. The neural network provides interpolation of the intensity pattern as well as the peak.

A comparison of how the neural network performs with CPS and PC inputs indicates that the method of input PC converges to a higher error than when using the CPS input and in order to minimize the error longer training times are required. The PC preprocessing computation also takes over four times as long as the CPS computation, at 23 FPS. When training the neural network, the model with the CPS-input takes converges to an initial error of 32% while the PC input converges to an initial error of 45%. The CPS model, when sequentially trained converges to 5% error as shown in [Fig. 2.3 (D)]. The PC includes the A value, which designates the scaled delta function and presents information relating to the translation between two similar images. By simply normalizing both the x-axis translation of A from the center of the PC and the gyroscope position for the x-position, the position estimated by A follows a trend similar to the actual position values.

Chapter 3

Discussion and Conclusions

Unlike 2D motion, which is easily inferred via standard phase correlation approaches and downward facing cameras, the 3D motion from a forward-directed camera is neither an analytic relation of the CPS, nor easily obtained from the CPS. Using a forward-facing camera is relevant in tasks where objects or obstacles in the flightpath of the UAV need to be avoided. Therefore, we focus on Fourier-domain PC methods with images captured by forward-facing cameras to infer 3D motion. Although not explicitly defined, we find the relative 3-D position is extractable from the CPS with regression-based, shallow neural networks.

Hybrid systems, which combine optical preprocessors with shallow neural networks should enable faster task-based image processing without requiring deep learning.

Event cameras (with temporal-stream processors [12]) or sensor-based processors (with additional pooling or Fourier-domain processing [31, 32]) may advance Fourier-domain processing with the simplest of neural networks. While with perfect images and coherent light, it is optimal to use the PC, which enables the most compression for optical odometry, results in practice are different.

In our demonstration and experiments, we find that while numerically, it is easier to input the PC into the neural network in practice since a neural network is capable of decoding patterns that don't have fringes. Prior work has emphasized the advantage of robustness to noise with shallow, "small brain' neural networks [36].

The neural network model converges with a few thousand CPS training images and also converges faster when the PC is input into the neural network instead of the CPS. However, calculating the PC requires an extra step during preprocessing, adding to the limitations of the approach. By sequentially training the same model on different paths or datasets, the neural network error when tested on unseen data reduces. We achieve MAE for object tracking within a few percent, which can be smoothed with a low-pass filter to recreate flight paths or track an object.

Currently, our approach achieves real-time image processing rates. With optically-computed CPS and Fourier-domain optical processing, the model would achieve even high-speed video processing rates. Our approach has shown issues handling a large spatial disparity between sequential images and the method, as implemented, requires supervised training with a labeled data set. We have not begun to approach issues such as image segmentation that are needed for self-driving cars nor have we studied the robustness to noise or conducted any ablation analysis. However, our research opens the door to new applications of CPS for high-sensitivity gyroscopes.

In conclusion, we perform an experiment using sequential training, image Fourier preprocessing, and shallow neural networks for visual odometry and object tracking. In Experiment 1, we calculate the CPS from sequential images of video across the five flightpath datasets. In Experiment 2, we sample each Fourier-plane image between diffraction modes across 8 object-scattered datasets. Our sequential training approach appears to function as U-net to filter the appropriate spectral representation. Our sequential, CPS preprocessing, and shallow neural network approach is:

- trainable with smaller batches of trajectories
- low power and requires less energy to train
- generalizable; after each trajectory training, the model demonstrates higher accuracy on an unseen trajectory or position
- uncertainty-limited; the cropped CPS provides velocity cut-offs and interpretable connections between inputs
- capable of low-latency, optical hybrid imaging processing with coherent light
- capable of high inference rates, 22k-52k FPS

In our study with sequential training, we achieve coarse generalizability with modest error (3-5%). While it remains an open question whether an entire scene may be reconstructed with shallow neural networks, we believe there is significant potential for real-time visual odometry or 3-D object tracking applications with our approach. If the CPS preprocessing were to be carried out optically, the inference speed can be increaesed further, as well as implemented on even more simple systems. Such high-speed inference systems may be particularly well-suited for obstacle avoidance in low size, weight, and power applications that need to function real-time in GPS-denied environments.

Bibliography

[1] C.D. Kuglin and D.C. Hines. The phase correlation image alignment method. *Proc. of Intl. Conf. on Cybernetics and Society*, 86(11):163–165, 1975.

[2] M.H.G. Peeters. *Implementation of the phase correlation algorithm : motion estimation in the frequency domain.* TU Eindhoven. Fac. Elektrotechniek : stageverslagen. Technische Universiteit Eindhoven, 2003.

[3] S UDOMKESMALEE. Shape-recognition in the fourier domain. In B JAVIDI, editor, *OPTICAL INFORMATION PROCESSING SYSTEMS AND ARCHITECTURES III*, volume 1564 of *PROCEEDINGS OF THE SOCIETY OF PHOTO-OPTICAL INSTRUMENTATION ENGINEERS (SPIE)*, pages 464–473. SOC PHOTO OPT INSTRUMENTAT ENGINEERS, 1991. CONF ON OPTICAL INFORMATION PROCESSING SYSTEMS AND ARCHITECTURES 3, SAN DIEGO, CA, JUL 23-26, 1991.

[4] Kaaren May and Nicholas Krouglicof. Moving target detection for sense and avoid using regional phase correlation. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, May 2013.

[5] Xue Wan, Jianguo Liu, Hongshi Yan, and Gareth L.K. Morgan. Illumination-invariant image matching for autonomous UAV localisation based on optical sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119:198–213, September 2016.

[6] R. Gonzalez. Improving phase correlation for image registration. In *IVCNZ 2011*, 2011.

[7] H. Foroosh, J.B. Zerubia, and M. Berthod. Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, 11(3):188–200, March 2002.

[8] Jocival Dias Junior, André Backes, Maurício Escarpinati, Leandro Silva, Breno Costa, and Marcelo Avelar. Assessing the adequability of FFT-based methods on registration of UAV-multispectral images. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2020.

[9] Chen Wang, Tete Ji, Thien-Minh Nguyen, and Lihua Xie. Correlation flow: Robust optical flow using kernel cross-correlators. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2018.

[10] B.S. Reddy and B.N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, August 1996.

[11] Xiaohua Tong, Kuifeng Luan, Uwe Stilla, Zhen Ye, Yusheng Xu, Sa Gao, Huan Xie, Qian Du, Shijie Liu, Xiong Xu, and Sicong Liu. Image registration with fourier-based image correlation: A comprehensive review of developments and applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(10):4062–4081, October 2019.

[12] M. Grédiac, F. Sur, and B. Blaysat. The grid method for in-plane displacement and strain measurement: A review and analysis. *Strain*, 52(3):205–243, May 2016.

[13] Harold S Stone, Bo Tao, and Morgan McGuire. Analysis of image registration noise due to rotationally dependent aliasing. *Journal of Visual Communication and Image Representation*, 14(2):114–135, 2003.

[14] G Tzimiropoulos, V Argyriou, S Zafeiriou, and T Stathaki. Robust FFT-based scale-invariant image registration with image gradients. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1899–1906, October 2010.

[15] Mohammad O. A. Aqel, Mohammad H. Marhaban, M. Iqbal Saripan, and Napsiah Bt. Ismail. Review of visual odometry: types, approaches, challenges, and applications. *SpringerPlus*, 5(1), October 2016.

[16] Hsiu-Wen Cheng, Tsung-Lin Chen, and Chung-Hao Tien. Motion estimation by hybrid optical flow technology for UAV landing in an unvisited area. *Sensors*, 19(6):1380, March 2019.

[17] Hsiu-Wen Cheng, Tsung-Lin Chen, and Chung-Hao Tien. Learning-based risk assessment and motion estimation by vision for unmanned aerial vehicle landing in an unvisited area. *Journal of Electronic Imaging*, 28(06):1, December 2019.

[18] Adric Eckstein and Pavlos P Vlachos. Digital particle image velocimetry (DPIV) robust phase correlation. *Measurement Science and Technology*, 20(5):055401, April 2009.

[19] R.R Meyer, A.I Kirkland, and W.O Saxton. A new method for the determination of the wave aberration function for high resolution TEM. *Ultramicroscopy*, 92(2):89–109, July 2002.

[20] TAKAO SUZUKI and TIM COLONIUS. Instability waves in a subsonic round jet detected using a near-field phased microphone array. *Journal of Fluid Mechanics*, 565:197, September 2006.

[21] C. A. Glasbey and K. V. Mardia. A penalized likelihood approach to image warping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):465–492, August 2001.

[22] Merwan Birem, Richard Kleihorst, and Norddin El-Ghouti. Visual odometry based on the fourier transform using a monocular ground-facing camera. *Journal of Real-Time Image Processing*, 14(3):637–646, July 2017.

[23] S. Wang, R. Clark, H. Wen, and N. Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050, 2017.

[24] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian M. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018.

[25] George Barbastathis, Aydogan Ozcan, and Guohai Situ. On the use of deep learning for computational imaging. *Optica*, 6(8):921, July 2019.

[26] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2014.

[27] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, April 2017.

[28] Jawad N. Yasin, Sherif A. S. Mohamed, Mohammad-Hashem Haghbayan, Jukka Heikkonen, Hannu Tenhunen, and Juha Plosila. Unmanned aerial vehicles (UAVs): Collision avoidance systems and approaches. *IEEE Access*, 8:105139–105155, 2020.

[29] Zhixiang Min, Yiding Yang, and Enrique Dunn. Voldor: Visual odometry from log-logistic dense optical flow residuals. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4897–4908, 2020.

[30] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong li. Displacement-invariant matching cost learning for accurate optical flow estimation, 10 2020.

[31] Ákos Zarándy, editor. *Focal-Plane Sensor-Processor Chips*. Springer New York, 2011.

[32] Riku Murai, Sajad Saeedi, and Paul H. J. Kelly. Bit-vo: Visual odometry at 300 fps using binary features from the focal plane, 2020.

[33] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[34] Seung-Hwan Baek and Felix Heide. Polka lines: Learning structured illumination and reconstruction for active stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5757–5767, June 2021.

[35] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific Reports*, 8(1), August 2018.

[36] Baurzhan Muminov and Luat T. Vuong. Fourier optical preprocessing in lieu of deep learning. *Optica*, 7(9):1079, August 2020.

[37] Tao Yan, Jiamin Wu, Tiankuang Zhou, Hao Xie, Feng Xu, Jingtao Fan, Lu Fang, Xing Lin, and Qionghai Dai. Fourier-space diffractive deep neural network. *Physical Review Letters*, 123(2), July 2019.

[38] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A. B. Miller, and Demetri Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588(7836):39–47, December 2020.

[39] Ravi Athale and Demetri Psaltis. Optical computing: Past and future. *Optics and Photonics News*, 27(6):32, June 2016.

[40] Mario Miscuglio, Zibo Hu, Shurui Li, Jonathan K. George, Roberto Capanna, Hamed Dalir, Philippe M. Bardet, Puneet Gupta, and Volker J. Sorger. Massively parallel amplitude-only fourier neural network. *Optica*, 7(12):1812, December 2020.

[41] Zhicheng Wu, Ming Zhou, Erfan Khoram, Boyuan Liu, and Zongfu Yu. Neuromorphic metasurface. *Photonics Research*, 8(1):46, December 2019.

[42] Erfan Khoram, Ang Chen, Dianjing Liu, Lei Ying, Qiqi Wang, Ming Yuan, and Zongfu Yu. Nanophotonic media for artificial neural inference. *Photonics Research*, 7(8):823, July 2019.

[43] Ying Zuo, Bohan Li, Yujun Zhao, Yue Jiang, You-Chiuan Chen, Peng Chen, Gyu-Boong Jo, Junwei Liu, and Shengwang Du. All-optical neural network with nonlinear activation functions. *Optica*, 6(9):1132, August 2019.

[44] Seung-Hwan Baek, Hayato Ikoma, Daniel S. Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H. Kim. End-to-end hyperspectral-depth imaging with learned diffractive optics, 2020.

[45] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[46] Oren Rippel, Jasper Snoek, and Ryan Adams. Spectral representations for convolutional neural networks. *NIPs*, 06 2015.

[47] Huan N. Do, Jongeun Choi, Chae Young Lim, and Tapabrata Maiti. Appearance-based localization of mobile robots using group LASSO regression. *Journal of Dynamic Systems, Measurement, and Control*, 140(9), April 2018.

[48] R. Carrillo Mendoza, P. Vera Bustamante, Brian, E. Hernández Castillo, and J. M. Ibarra Zannatha. 3d self-localization for humanoid robots using view regression and odometry. In *2015 12th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–5, 2015.

[49] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, 5(40), March 2020.

[50] Amado Antonini, Winter Guerra, Varun Murali, Thomas Sayre-McCord, and Sertac Karaman. The blackbird UAV dataset. *The International Journal of Robotics Research*, 39(10-11):1346–1364, March 2020.

[51] William T. Rhodes. titlehistory and evolution of the teaching of fourier optics/title. In Angela M. Guzman, editor, *3rd Iberoamerican Optics Meeting and 6th Latin American Meeting on Optics, Lasers, and Their Applications*. SPIE, July 1999.

[52] Adric C. Eckstein, John Charonko, and Pavlos Vlachos. Phase correlation processing for DPIV measurements. *Experiments in Fluids*, 45(3):485–500, March 2008.

[53] Thomas Schanze and Reinhard Eckhorn. Phase correlation of cortical rhythms at different frequencies: higher-order spectral analysis of multiple-microelectrode recordings from cat and monkey visual cortex. *International Journal of Psychophysiology*, 26(1-3):171–189, June 1997.

[54] A. M. Cormack. Fourier transforms in cylindrical coordinates. *Acta Crystallographica*, 10(5):354–358, May 1957.

[55] Peter Muller and Andreas Savakis. Flowdometry: An optical flow and deep learning based approach to visual odometry. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2017.

[56] Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral DiffuserCam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10):1298, September 2020.

[57] Joseph W Goodman. Introduction to fourier optics. *Introduction to Fourier optics, 1st ed., by JW Goodman. McGraw-Hill*, 1, 1068.

[58] David Casasent and Demetri Psaltis. Position, rotation, and scale invariant optical correlation. *Applied Optics*, 15(7):1795, July 1976.

[59] Stéphane Derrode and Faouzi Ghorbel. Robust and efficient fourier–mellin transform approximations for gray-level image reconstruction and complete invariant description. *Computer Vision and Image Understanding*, 83(1):57–78, July 2001.

[60] Ming Fang and Gerd Hausler. Class of transforms invariant under shift, rotation, and scaling. *Applied Optics*, 29(5):704, February 1990.

[61] Mehjabin Sultana Monjur, Shih Tseng, Renu Tripathi, and M. S. Shahriar. Incorporation of polar mellin transform in a hybrid optoelectronic correlator for scale and rotation invariant target recognition. *Journal of the Optical Society of America A*, 31(6):1259, May 2014.

[62] Demetri Psaltis and David Casasent. Deformation invariant optical processors using coordinate transformations. *Applied Optics*, 16(8):2288, August 1977.

[63] Toyohiko Yatagai, Kazuhiko Choji, and Hiroyoshi Saito. Pattern classification using optical mellin transform and circular photodiode array. *Optics Communications*, 38(3):162–165, August 1981.

[64] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. DeepSat. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*. ACM Press, 2015.

[65] Jianguo Zhang and Tieniu Tan. Brief review of invariant texture analysis methods. *Pattern Recognition*, 35(3):735–747, March 2002.

[66] Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljačić. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441–446, June 2017.

[67] Matthew Hutson. AI researchers allege that machine learning is alchemy. *Science*, May 2018.

[68] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, July 2018.

[69] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[70] Luat Vuong and Hobson Lane. Nonlinear spectral preprocessing for small-brain machine learning. In Michael E. Zelinski, Tarek M. Taha, Jonathan Howe, Abdul A. Awwal, and Khan M. Iftekharuddin, editors, *Applications of Machine Learning*. SPIE, September 2019.

[71] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):1–12, November 2016.

[72] Lasitha Piyathilaka and Rohan Munasinghe. An experimental study on using visual odometry for short-run self localization of field robot. In *2010 Fifth International Conference on Information and Automation for Sustainability*, pages 150–155. IEEE, 2010.

[73] Tawfiq A Al-Assadi and Abdulkadhem Abdulkareem Abdulkadhem. Trajectory extraction, simplification and representation based on phase correlation and dynamic chain code for forward facing camera. *Journal of Computational and Theoretical Nanoscience*, 16(3):861–868, 2019.