Insect Identification in Pulsed Lidar Images Using Changepoint Detection Algorithms

Nathaniel Sweeney[1], Caroline Xu[2], Joseph A. Shaw[1,3], Toby D. Hocking[4], and Bradley M. Whitaker[1,3]
[1]*Electrical and Computer Engineering, Montana State University, USA*
[2]*Electrical Engineering and Computer Science, University of Michigan, USA*
[3]*Optical Technology Center, Montana State University, USA*
[4]*School of Informatics, Computing, and Cyber Systems, Northern Arizona University, USA*
bradley.whitaker1@montana.edu

***Abstract*—Noninvasive entomological insect monitoring often utilizes a variety of tools such as lidar to gather information without interfering with the insects in their habitat. These collection methods often result in large amounts of data that can be tedious and lengthy to interpret and analyze. Machine learning has been previously used in the past in order to analyze lidar images to detect insects, but often suffers from pitfalls such as long training times and large computational power requirements. In an attempt to offer an alternative that takes little to no training on the data and much less computational power, this paper looks at the use of changepoint detection algorithms to analyze lidar images containing insects. By analyzing the rows or columns of a lidar image, the algorithms should be able to detect abrupt changes in the image that would represent the insects. While not as accurate, the changepoint detection algorithms give comparable results to a machine learning algorithm tested on the same dataset without the need for supervised training.**

***Index Terms*—Lidar, Changepoint Analysis, Anomaly Detection**

I. Introduction

A. Background

Entomological methods for the monitoring of insects traditionally uses manual inspection techniques such as traps [1] or catching insects [2] in order to gather insects. These methods are often labor intensive, time-consuming, and disruptive to the insects, thus leading to the use of noninvasive insect monitoring with tools such as lidar [3]. These noninvasive methods can capture entomological information without disrupting the insects in their habitat, while also reducing the amount of manual labor required for collecting insect data. Previous work has been done using lidar to identify disease-carrying mosquitoes [4], [5] and characterizing agricultural pests [6], [7]. Other pulsed lidar methods have been developed specifically to detect the insect wingbeat modulations [8], [9], [10]. However, much of the analysis of the information from these techniques often requires large amounts of time and effort to manually analyze the information. Previous work has been done with the use of machine learning algorithms to help automate the analysis of lidar images and cut down on the amount of time it takes to analyze the data [11], [12]. While these have been shown to be successful, some of the many shortcomings of machine learning are often long training times of algorithms and large amounts of computational power required for the algorithms.

As an alternative to machine learning, changepoint detection algorithms could be used to analyze the images while greatly cutting down on the time it takes and the computational power required. This paper shows work done on two different datasets of insect-containing lidar images using two different changepoint detection algorithms, a graph-constrained changepoint detection algorithm, gfpop [13], [14], and MATLAB's `findchangepts` function.

B. Changepoint Detection

Changepoint detection algorithms are developed to detect abrupt changes in a variety of values such as mean, variance, or standard deviation. They are often utilized in fields such as medicine, neuroscience or genomics where it is necessary to find abrupt changes in large data sequences.While there are a large number of changepoint detection algorithms, the two that were used in this paper are gfpop (Graph Constrained Functional Pruning Optimal Partitioning) [13], [14] and MATLAB's `findchangepts` function.

The first algorithm that was used was gfpop. This changepoint detection algorithm was proposed by Hocking et al. [13] and later implemented into an R interface by Runge et al. [14]. In gfpop, input data is analyzed based on of a user-implemented graph with nodes to represent the various states that the data could be in at any given point and penalized edges representing the transitions between the various states. The algorithm has a variety of loss functions that can be utilized as well as several different types of transition edges and other edge parameters such as the penalty, decay, threshold for biweight and Huber losses, and the slope for the Huber robust loss. While the work in this paper only utilized the penalty parameter, other applications are more likely to utilize these parameters. The reason for using this algorithm is due to the fact that the user could have a good idea of what the dataset should look like prior to collection, allowing them to quicky generate a graph and iterate through the dataset with little to no training and testing outside of tweaking the penalties on the edges.

MATLAB's `findchangepts` function is a part of the Signal Processing Toolbox and analyzes the input data to the function while returning the most significant changes in the

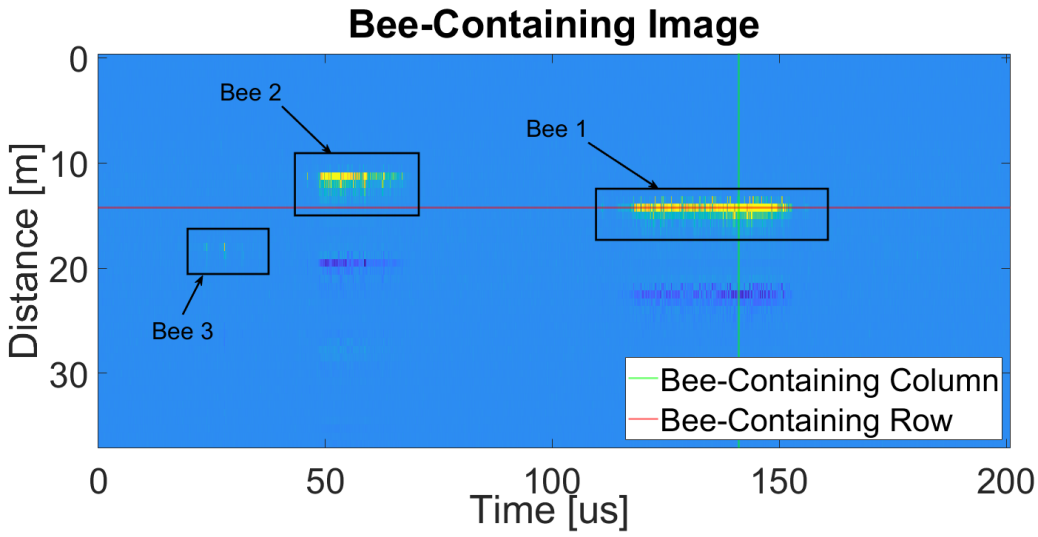

Fig. 1. Example lidar image containing three bees.

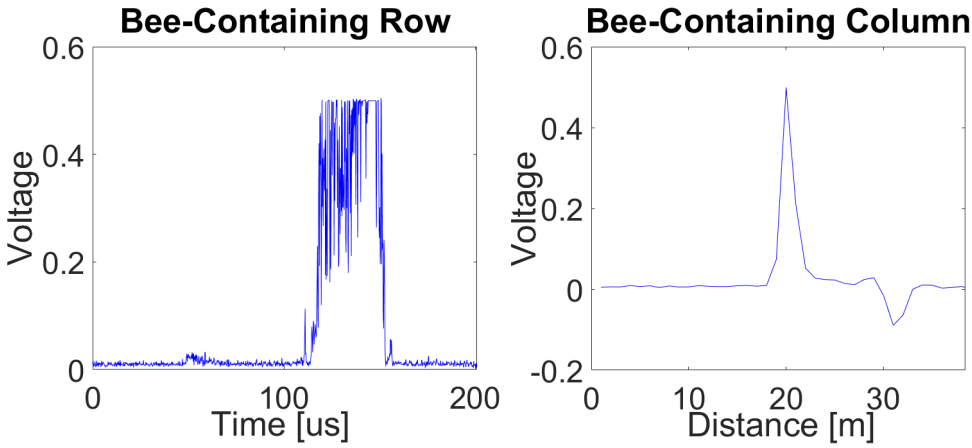

Fig. 2. Example of an insect row and column

data for several different types of statistics such as mean, RMS value, or standard deviation.

II. Methods

A. Datasets

1) Beehives Dataset: The first dataset used for both gfpop and `findchangepts` is a set of lidar images taken from beehives located at the Horticulture Farms of Montana State University in Bozeman during June 2022. The dataset contains 4131 images (178x1024), 1309 of which contain a bee based on labels manually generated by five different people over the course of a combined 35 hours. Each bee also has a confidence level, ranging from 1 (less confident) to 5 (most confident). In these images, the bee is often represented by a spike in energy if the laser hit their body, or a series of smaller spikes in energy if the laser hit the wings of the bee (see Fig. 1). In the case of the latter, the frequency of the flutters of energy can show the harmonics of the oscillating wings. Both types of spikes can be seen in both the rows and columns of the images (see Fig. 2).

2) Hyalite Creek Dataset: Both algorithms were further tested on previously taken lidar images of insects at Hyalite Creek outside of Bozeman, MT during September of 2020. This dataset contains 10,079 images (178x1024) that represent similar info to the beehives dataset. However, this dataset is far more sparse than the beehives dataset, with only 172 of the 10,079 images labeled as containing an insect, and with a large sum of those also labeled only as maybe an insect. Similarly to the beehives dataset, these images were manually labeled over a long period of time that isn't exactly known. Testing was done on this dataset in order to allow a direct comparison to machine learning algorithms that were that were tested on the dataset by Vannoy et al. [11].

B. Preprocessing

For the gfpop algorithm, the only preprocessing that was done involved shifting any of the negative values present in the image, mostly artifacts from the lidar return, to zero. Each row or column, depending on which the algorithm was going to be applied to, was then smoothed with a three point moving mean window to reduce the open air noise and the noise at the top of the rows containing an insect.

For the `findchangepts` algorithm, additional preprocessing was performed. All of the rows that were guaranteed to not contain an insect were removed. This was done by splitting the row into eight windows and finding the average of each. If all of these were less than 90% of the average of the whole image, it was considered an empty row. Hard targets were then removed in a similar fashion, this time checking if each of the averages of the windows was greater than 0.8. This value was chosen because the images were normalized from 0 to 1. An additional threshold of 2.25*[Image Average] was then applied in a similar fashion. Noisy rows were then removed by dividing rows into 16 windows and finding their average. The percent difference between each section was found and if the difference was within 5% then the row was considered insignificant and removed. A wavelet transform is then performed on the remaining rows using the `cwt` function in order to distinguish between the remaining rows that contain wings of insects and other targets. If the maximum value that is returned is less than 0.4, the row is considered unimportant. The remaining scalograms are then split into 8 sections and the average of each rectangle is found. If the percent different between them is less than 10%, then the row is removed. If there were no rows remaining in the image, it would be considered to have no insects.

C. gfpop

The main methodology of applying gfpop to the lidar images involved sending data from the images through the algorithm with the changepoint graph shown in Figure 3. Both the rows and the columns were sent through the algorithm to compare the results between the two, although the same graph and penalties were used. The selected graph was designed to allow gradual increases to the insect states, resulting in higher accuracy as opposed to a single jump to the insect state and then back down. To prevent the algorithms from continuously increasing the state value, the graph is also specified to start and end in the "air" state, even though some of the insects are present on the very edges of the dataset. Each state has a null edge on it which allows the dataset to remain in the state that it is in if the changes in the data aren't significant enough to trigger a state change. The increasing and decreasing states also both have an up edge transition and a down edge transition respectively, which is what lets the graph gradually increase to the peak, although this is mostly only relevant in the analysis of the rows since the insect columns are usually only three data points in the series. Lastly, the top of the graph which represents the actual insect has a penalty for decreasing back down from the peak.

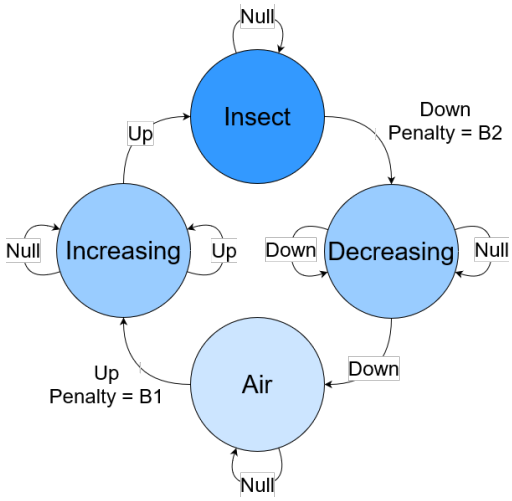

Fig. 3. gfpop graph for insect detection on both datasets.

After sending each row and column of the image through the algorithm, it saves the all of the changepoints, including their location, their state, and the actual data parameter of the changepoint. After getting results back from the algorithm, the parameter of any insect changepoints were checked on whether they were double the mean of the row where the changepoint was detected and 1.5 times the mean of the overall image. This helped to remove any false positives that can be attributed to noise in the air or hard targets that were present in any of the images. Because the insect data was normalized from 0 to 1, while the beehive dataset uses raw voltages, the penalty was set to 0.005 for the beehives dataset and 0.1 in the Hyalite Creek dataset.

D. MATLAB's `findchangepts`

For MATLAB's `findchangepts`, an algorithm using the function was developed for both the bees and insects datasets separately, mainly due to the fact that the insect dataset was far more sparse than the bee dataset. For the insect algorithm, the remaining rows from preprocessing were all put into the `cwt` function, and the absolute value was calculated for the 71x1024 matrix that resulted from it, along with the maximum value of the matrix. Since peaks in the dataset are desired, any value in the matrix below a threshold (.225 times the maximum value) was set to 0, and the matrix was cropped to contain only the 40 top rows (other rows correspond to lower frequencies that do not have wingbeat modulation data) and normalized between 0 and 1. The average of the remaining columns was stored into a vector and the matrix was then smoothed and normalized again. This final vector was then sent through the `findchangepts` function. This analyzes the mean of the input with a minimum threshold of 3.9. The original signal is then smoothed and normalized and sent through the same function with a minimum threshold of 3. If no changes are found in this second result, there are no insects present in that row. If changepoints are found, then the changepoints of

the scalogram and the original image are compared to each other. If they are within 21 pixels of each other, roughly a 2% difference, then the scalogram changepoint is kept, otherwise it is discarded. After this, the location of the insect is found by finding the brightest points in the scalogram, however this process only accommodates for two insects per row. The locations of these insects are then saved, although exact locations weren't verified, only the overall images. After these locations are found, they are discarded if they are over 605 pixels long or less than 15 pixels long. The remaining changes are considered insects.

The bee algorithm is similar, with the same preprocessing but with reduced thresholds since there were significantly more changes in the bee dataset. The first changepoint threshold was reduced from 3.9 to 2 and the second was reduced from 3 to .8. The maximum length that a bee signal could be was also increased from 605 pixels to 700. These rows were then grouped by insect and the middle column of the signal was found. It was then normalized and put into the `islocalmax` function with a minimum prominence of .6. If a peak was located at a row containing an insect signal it was considered an insect.

III. Results

The results of both algorithms were fairly comparable to one another, and can also be compared to some supervised machine learning results in the case of the Hyalite Creek insect dataset. The beehive results cannot be compared to any machine learning results due to the fact that no results from this dataset have been published. For the MATLAB algorithm, the results were found with a runtime of of 8066 seconds, resulting in a 73.3% accuracy. Of the 1309 insect-containing images, the algorithm detected 857 of them (65.5%). The algorithm also resulted in 651 false positives. For the gfpop algorithm on the rows, the results were compiled with a runtime of 1733 seconds, resulting in an accuracy of 76.11%. It was able to detect 1077 of the insect images, or 82.28%. There were 755 false positives. When gfpop was applied to the columns, it took a runtime of 2721 seconds, resulting in an accuracy of 90.11% and correctly identifying 1059 of the insect images, or 80.9%. It also resulted in 158 false positives.

For the Hyalite Creek dataset, the MATLAB algorithm was able to identify 122 of the 172 insect-containing images, or 71.1%, with a runtime of 14,474 seconds. It also incorrectly identified 196 images as insect-containing. For the gfpop algorithm on the rows, it properly identified 140 of the insect-containing images, or 81.4%, in a runtime of 4009 seconds with 1137 false positives. When it was applied to the columns, it found 106 of the insect-containing images with a runtime of 5906 seconds and 1840 false positives. In the case of the machine learning algorithms, specifically a neural network, the algorithm identified 32 of the 44 insect images that were in the testing set, or 72.7% respectively. When going over these results, accuracy was not considered due to the fact that the dataset was so sparse that simply guessing that there are no insects would result in an accuracy of 98%.

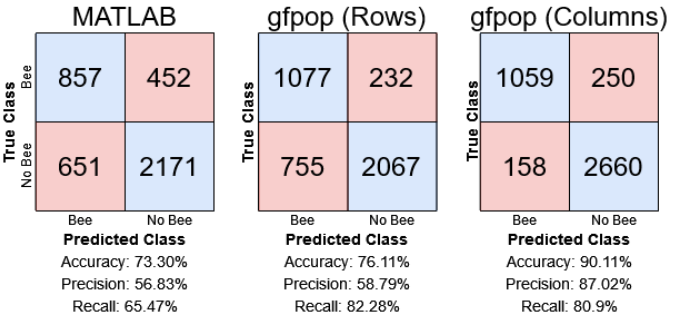

Fig. 4. Results from the beehives dataset.

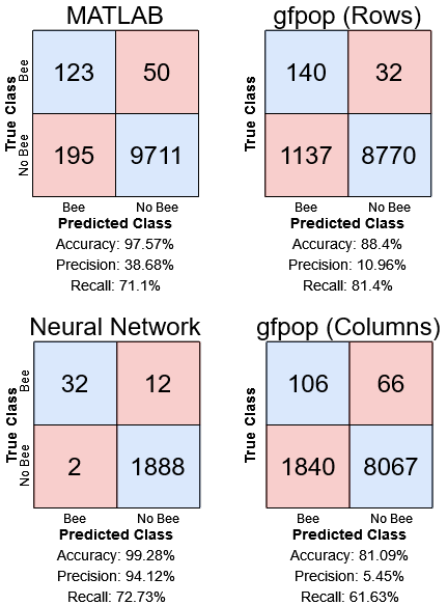

Fig. 5. Results from the Hyalite Creek dataset.

IV. Conclusions and Future Work

The results from the two algorithms were quite promising and were able to accomplish the main goals of reducing computational power and reduce the time it takes to analyze the data, while also generating comparable results to the machine learning algorithms. Compared to the 35 hours that it took to manually label the beehive dataset, and an even larger amount of time to label the Hyalite Creek dataset, the changepoint algorithms were able to cut down on the analysis time greatly. Further analysis of the results could also be done to count the specific number of insects that the algorithms were capable of finding as opposed to just the images that contain an insect. This could provide even more important information for entomological purposes.

Future work for this will focus on improving the results and attempting to use the algorithms as real-time detection methods when taking lidar images. For gfpop, further testing will involve making changes to the graphs and penalties that are used, while also applying more in depth post-processing

to verify any of the rows or columns that were marked as containing a bee. For `findchangepts`, future work would involve improving preprocessing on the images and adjusting thresholds for the function. The algorithm can also benefit from using parallel processing in order to decrease the runtime of the algorithm as opposed to running it linearly as it currently does. The algorithm could also be tested on the voltage matrices of the lidar images as opposed to the normalized data. This would result in more consistent values between each image for the air, hard targets, and bees. Since both algorithms are also implemented in MATLAB, testing could be done to use them as a real-time detection algorithm when taking lidar images. Future work will also consider developing penaly learning algorithms, rather than using fixed penalties, with gfpop [13].

ACKNOWLEDGMENT

This work was funded by the Air Force Research Laboratory (AFRL) on a subcontract from S2 Corporation under prime award No. FA8650-16-C-1954, with a fundamental research exemption.

REFERENCES

[1] H. Kawada, S. Honda, and M. Takagi, "Comparative Laboratory Study on the Reaction of Aedes aegypti and Aedes albopictus to Different Attractive Cues in a Mosquito Trap ," *Journal of Medical Entomology*, vol. 44, pp. 427–432, 05 2007.

[2] N. L. Achee, L. Youngblood, M. J. Bangs, J. V. Lavery, and S. James, "Considerations for the use of human participants in vector biology research: A tool for investigators and regulators," *Vector-Borne and Zoonotic Diseases*, vol. 15, no. 2, pp. 89–102, 2015. PMID: 25700039.

[3] M. Dwivedi, M. H. Shadab, and V. R. Santosh, *Insect Pest Detection, Migration and Monitoring Using Radar and LiDAR Systems*, pp. 61–76. Singapore: Springer Singapore, 2020.

[4] G. E. Batista, Y. Hao, E. Keogh, and A. Mafra-Neto, "Towards automatic classification on flying insects using inexpensive sensors," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 1, pp. 364–369, 2011.

[5] S. Jansson, P. Atkinson, R. Ignell, and M. Brydegaard, "First polarimetric investigation of malaria mosquitoes as lidar targets," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–8, 2019.

[6] Y. Y. Li, H. Zhang, Z. Duan, M. Lian, G. Y. Zhao, X. H. Sun, J. D. Hu, L. N. Gao, H. Q. Feng, and S. Svanberg, "Optical characterization of agricultural pest insects: a methodological study in the spectral and time domains," *Applied Physics B*, vol. 122, no. 8, p. 213, 2016.

[7] L. Mei, Z. G. Guan, H. J. Zhou, J. Lv, Z. R. Zhu, J. A. Cheng, F. J. Chen, C. Lofstedt, S. Svanberg, and G. Somesfalean, "Agricultural pest monitoring using fluorescence lidar techniques," *Applied Physics B*, vol. 106, no. 3, pp. 733–740, 2011.

[8] J. A. Shaw, N. L. Seldomridge, D. L. Dunkle, P. W. Nugent, L. H. Spangler, J. J. Bromenshenk, C. B. Henderson, J. H. Churnside, and J. J. Wilson, "Polarization lidar measurements of honey bees in flight for locating land mines," *Optics express*, vol. 13, no. 15, pp. 5853–5863, 2005.

[9] K. S. Repasky, J. A. Shaw, R. Scheppele, C. Melton, J. L. Carsten, and L. H. Spangler, "Optical detection of honeybees by use of wing-beat modulation of scattered laser light for locating explosives and land mines," *Applied optics*, vol. 45, no. 8, pp. 1839–1843, 2006.

[10] J. A. Shaw, K. S. Repasky, J. L. Carlsten, L. H. Spangler, and D. S. Hoffman, "Optical detection of oscillating targets using modulation of scattered laser light," Mar. 31 2009. US Patent 7,511,624.

[11] T. C. Vannoy, T. P. Scofield, J. A. Shaw, R. D. Logan, B. M. Whitaker, and E. M. Rehbein, "Detection of insects in class-imbalanced lidar field measurements," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021.

[12] H. B. Madsen, T. C. Vannoy, E. M. Rehbein, R. D. Logan, J. A. Shaw, and B. M. Whitaker, "Automated insect recognition in unlabeled lidar field measurements," in *Laser Radar Technology and Applications XXVII* (G. W. Kamerman, L. A. Magruder, and M. D. Turner, eds.), vol. 12110, p. 1211007, International Society for Optics and Photonics, SPIE, 2022.

[13] T. D. Hocking, G. Rigaill, P. Fearnhead, and G. Bourque, "Constrained dynamic programming and supervised penalty learning algorithms for peak detection in genomic data," *J. Mach. Learn. Res.*, vol. 21, jan 2020.

[14] V. Runge, T. D. Hocking, G. Romano, F. Afghah, P. Fearnhead, and G. Rigaill, "gfpop: an R package for univariate graph-constrained change-point detection," 2020.