Deep Learning and Fish Tags

Santosh Medisetty

Dave Ouellette

Frank Smith

Matthew Richard

Sam Johnston

Researchers explore the use of modern machine learning methods to automate the analysis of data from acoustic tags.

Who should read this paper?

Have you ever had to manually analyze hours upon hours of fish tracking data and found yourself wondering if a machine could be trained to solve the task for you? Then this paper is for you. You will also find this paper worth reading if you are curious about what is stirring in the world of acoustic fish tracking technology or interested in novel and surprising applications of machine learning.

Why is it important?

There is a growing interest in studying smaller fish and understanding their interactions with humans and the effects of modifications to fish habitats. A fish tracking technology has been invented suitable for tracking large numbers of fish simultaneously in noisy environments. However, the analysis of the receiver data is time-consuming and expensive to complete because some steps have to be done manually. The machine learning solution proposed in this paper has the potential to alleviate the need for manual validation, greatly accelerating the data processing.

The researchers have taken a deep neural network architecture known as U-Net, commonly used to solve image segmentation problems (e.g., analysis of MRI images or geospatial satellite imagery), and trained it to solve an entirely different task; namely, the U-Net has been trained to detect and distinguish acoustic tags based on their transmission rate. Further work is required before the machine learning solution can replace manual validation. But the researchers hope that the technology will be ready for commercial application within one year.

About the authors

Santosh Medisetty is a M.Sc. student in the Faculty of Computer Science at Dalhousie University.

Dave Ouellette is the senior software developer and analyst for fish tracking at Innovasea.

Frank Smith is the director of software and data products for fish tracking at Innovasea.

Matthew Richard is the machine learning engineer at Innovasea.

Sam Johnston is the principal investigator for fish tracking at Innovasea.

Jean Quirion is the vice president of research and development for fish tracking at Innovasea.

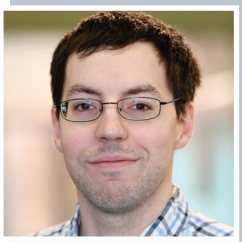
Dr. Jason Newport is the data manager at DeepSense.

Dr. Christopher Whidden, formerly with DeepSense, is now an assistant professor in the Faculty of Computer Science at Dalhousie University.

Dr. Oliver Kirsebom is the senior staff scientist at MERIDIAN and an adjunct professor in the Faculty of Computer Science at Dalhousie University.

Jean Quirion

Dr. Jason Newport



Dr. Christopher Whidden

Dr. Oliver Kirsebom

IDENTIFICATION OF PERIODIC FISH TAGS WITH DEEP LEARNING

Santosh Medisetty[1], Dave Ouellette[2], Frank Smith[2], Matthew Richard[2], Sam Johnston[2], Jean Quirion[2], Jason Newport[3], Christopher Whidden[1], and Oliver Kirsebom[1,4]
[1]Faculty of Computer Science, Dalhousie University, Halifax, N.S., Canada
[2]Innovasea, Halifax, N.S., Canada
[3]DeepSense, Dalhousie University, Halifax, N.S., Canada
[4]MERIDIAN, Dalhousie University, Halifax, N.S., Canada

ABSTRACT

Acoustic tags are increasingly being used in research to track movements of marine animals, generating large, complex datasets which require automated analysis methods. In this work, we explore the use of modern machine learning methods to automate the analysis of data from acoustic tags with a novel encoding developed for the study of large numbers of fish in noisy environments. These tags emit pulses at a configurable interval but without any other encoded identification message. This reduces power usage and enables continuous tracking but the challenge is that tags must be identified based on the transmission rate and the position reconstructed by triangulating detections by at least three independent receivers. Utilizing a visual representation of the time series data generated by the tag, we adapt a deep neural network architecture known as U-Net to the task of identifying individual acoustic tag transmissions. Testing our model on unseen data, we achieve an accuracy of 97%, outperforming the solution currently in use, which achieves an accuracy of 41%. Our model shows significant promise and represents the first step towards enabling the large-scale application of this new acoustic tracking technology to fisheries operations.

INTRODUCTION

Tracking the movements of marine animals in space and time plays an increasingly important role in the sustainable management of ocean resources [Hussey et al., 2015]. In the context of fishery operations research, there is a desire to utilize fish tracking technology to gain a better understanding of fish movement, allowing fisheries to optimize both operations and sustainability by improving the targeting of desired fish stocks while reducing bycatch.

In particular, studies that require large numbers of tags in one place at one time are increasing in numbers and importance as we study smaller fish and their interactions with human modifications to their habitats [Muñoz et al., 2020; Aspillaga et al., 2021]. Unfortunately, commonly employed acoustic tags, which transmit complex sequences of acoustic pulses for identification, are not well suited for such studies because they are susceptible to interference from other tags, multi-path signals, and noise signals. Innovasea, an ocean technology company headquartered in Halifax, N.S., has invented an improved fish tracking solution that employs a novel encoding scheme [Ehrenberg and Steig, 2003; 2009] in which tags are identified based on the transmission rate as opposed to a complex pulse sequence. Past studies have shown that this coding scheme can provide high resolution positioning and tracking of tagged fish for large numbers of fish simultaneously in noisy and varied environments and at farther ranges [Steig et al., 2004; Ransom et al., 2008; Semmens, 2008]. Therefore, it has the potential to enable improved tracking for large numbers of tagged fish in noisy aquatic environments. The results of these studies came at a high cost, however – the analysis of the receiver data was time-consuming and expensive to complete because some steps had to be done manually. This creates a bottleneck in the data processing which has made it difficult to use the solution in fishery operations.

In this paper, we discuss how deep learning techniques can help automate the analysis of the data generated by the novel encoding scheme, paving the way for applications to fishery operations. Within the last decade, deep neural networks have become the preferred machine learning approach for solving a wide range of tasks, outperforming existing computational methods and achieving human-level accuracy in domains such as image classification [He et al., 2015] and natural speech processing [Hinton et al., 2012]. Originally inspired by the human brain, neural networks consist of a large number of interconnected "neurons," each typically performing a simple linear operation on input data, specified by a set of weights and a bias, followed by an activation function. In a supervised training approach, the network is given examples of labelled data, and the weights and biases are adjusted to produce the desired output using an optimization algorithm. Modern neural networks exhibit multi-layer architectures, which enable them to build complex concepts out of simpler concepts and, hence, learn a non-linear representation of the data conducive to solving a given task. One of the most commonly encountered architectures is the convolutional neural network (CNN) which is particularly well adapted to the tasks of analyzing image data owing to its use of weight-sharing filters that slide across the image [Goodfellow et al., 2016].

We adapt a particular variation of the common CNN known as U-Net [Ronneberger et al., 2015], developed specifically for solving image segmentation tasks, to analyze an image representation of the acoustic time series data. We identify several scenarios in which the analysis becomes particularly challenging and devise data augmentation strategies to overcome these challenges. Finally, we test our U-Net model on unseen tag data achieving near-human accuracy and outperforming the only existing automated solution by a significant margin.

DATA COLLECTION AND PROCESSING

Encoding Scheme

In the encoding scheme [Ehrenberg and Steig, 2003; 2009], the fish tags transmit short-duration acoustic pulses, referred to as "pings," at regular intervals that are detected by dedicated underwater receivers. The tags are then identified from the observed delay between successive pings. In typical applications, the delay is between 1 s and 10 s, where shorter periods yield higher spatial resolution, but also shorten the tag's lifetime due to increased energy consumption. This tradeoff between resolution and energy consumption drives the need for more efficient encoding schemes. Moreover, the tags can be configured to emit not one, but two, closely spaced pings. Varying the temporal separation of the two pings gives another way to distinguish between tags.

The strength of the encoding scheme lies in its simplicity. Other commonly used encoding schemes employ complex sequences of pulses. Such schemes allow tags to be uniquely identified based on a single transmission but consume more energy and are more susceptible to noise interference. As a direct result, the novel encoding scheme has the potential to provide higher tracking resolution and enhanced range.

Image Representation

While the new encoding scheme sounds simple to implement, in practice it can be rather challenging to identify acoustic tags based on their ping rate, especially if multiple transmitting tags are within the detection range of the receiver simultaneously. Moreover, natural environment noise signals and multi-path reflections of pings clutter the picture and must be ignored. To facilitate the identification of tags based on their ping rate, an image representation has been developed in which the received pings are plotted according to their time of arrival (x) and their displacement (y) with respect to a chosen clock rate. In this representation, pings originating from a (stationary) tag with a ping rate that matches the clock rate will describe a horizontal track, allowing them to be identified and "marked" by a human analyst through visual inspection. As a tagged fish moves, the track will deviate slightly from horizontal due to the Doppler effect, conveying important information about the fish's motion, but also complicating the marking task.

In Figure 1, we show an example of the image representation used to identify the acoustic tags. Note that for improved visibility only part of the y-axis is shown; the full y-axis extends from 0 s to a little over 9 s (the period of the acoustic tag). The raw pings are shown as black dots, while the pixels identified

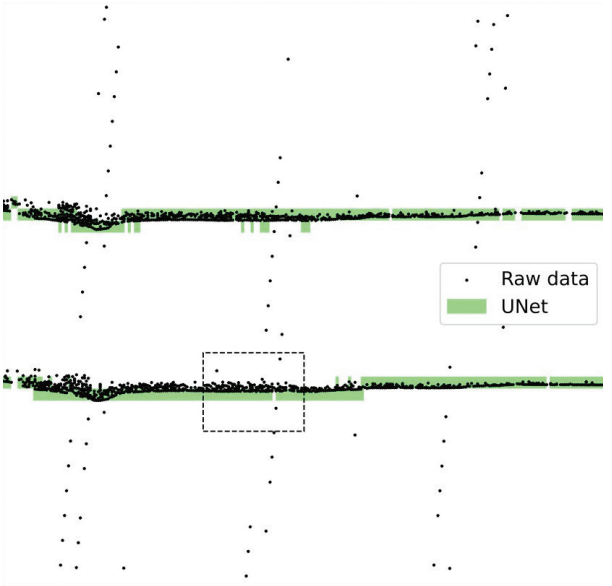


Figure 1: Scatter plot showing an example of the image representation used to identify the acoustic tags. Note that for improved visibility only part of the y-axis is shown; the full y-axis extends from 0 s to a little over 9 s. The black dots are the raw pings while the shaded green regions are the pixels identified by the U-Net machine learning solution proposed in this work. A zoom-in of the region enclosed by the box (dashed line) is shown in Figure 4.

by our proposed U-Net machine-learning solution (discussed below) are shown as the shaded (green) area.

In a manual review, the human analyst typically inspects the images at varying temporal scales and resolutions to better identify any track-like features. However, for our first attempt at automating the image analysis task, we opted to work with a single temporal scale and resolution to simplify the model development. Specifically, we used an image size of 192 x 192 bins with an x-axis (temporal) range of 30 minutes, while the y-axis range was adjusted to match the known tag period which were in the range of 9-10 s for the dataset we worked with. This implies an x-axis time resolution of 9.4 s and a y-axis time resolution of around 50 ms. While we did not perform a systematic investigation of the optimal image size and resolution, it seems plausible that an x-axis time resolution close to the tag period should provide optimal performance. (We note that the particular machine learning model used in this work requires the input image to have dimensions that are integer multiples of 16.)

The image representation described above requires prior knowledge of the tag period and, hence, the tag identification method proposed here depends on this information. In most application scenarios, this information will be available, but there are also scenarios in which it may be of interest to be able to detect acoustic tags without prior knowledge of the period, such as detecting tagged fish that have travelled long distances. Developing an automated tag identification solution that does not require prior knowledge of the period is an interesting problem, which, however, is beyond the scope of the present work. We discuss this problem further in the "Conclusion and Outlook" section.

Position Reconstruction

In order to reconstruct the position of an acoustic tag, its transmissions must be detected independently by at least three hydrophones. Since, as a general rule, acoustic tags become harder to detect the

further they are from the receiver, this has important implications for the design of hydrophone arrays, if one is interested in tracking the movements of fish across a site without gaps in coverage.

In the image representation discussed above, a distant acoustic tag will produce a faint track because a high proportion of the emitted pings do not reach the hydrophone (e.g., due to reflections of natural obstacles such as rocks) or are too weak to be detected by the time they reach the hydrophone (due to attenuation). Therefore, it is of interest to develop a marking solution that is able to detect as faint tracks as possible, as this will enable acoustic tags to be detected at farther ranges, allowing researchers to cover a given study area with fewer hydrophones, thereby reducing deployment costs significantly.

Existing Automated Solution – MarkTags

Currently a software solution known as MarkTags, which automates the marking of acoustic tags, is being utilized. MarkTags utilizes "conventional" rule-based algorithms rather than a machine learning approach, e.g., algorithms that remove noise by filtering based on the expected tag period or make assumptions about the expected trajectory of the tag. The MarkTags software has several parameters that can be adjusted to optimize its performance. However, for most applications the accuracy remains insufficient. To the best of our knowledge, MarkTags is the only existing, automated solution for identifying acoustic tags based solely on their transmission rate.

Dataset

For this work, we used a dataset originating from a study conducted in a shallow river environment in a location in northern California, U.S., employing an array of 10 hydrophones and 35 acoustic tags, 25 of which were attached to juvenile Green Sturgeon while 10 were time-synchronous tags used to synchronize the receivers with the GPS clock. The hydrophones were of the type HR3-307 kHz (Innovasea, N.S.) while the acoustics tags were of the type V3-307 kHz (Innovasea, N.S.). The tags weigh 0.3 g, measure 15 mm in length and 4 mm in diameter, and were programmed to transmit pings with periods in the range of 9-10 s. Data were recorded for 43 days and subject to manual marking, resulting in the identification of 21.5 million pings.

Data Structure

The dataset contains 426 raw data files and the same number of files with the results of the manual marking effort. Each file contains data from one day (24 hours) of recording from a single hydrophone. A small excerpt of a marked file is shown in Figure 2.

The data consists of a time series of detected pings. For each ping, the following variables are recorded as shown in Table 1. In this work, we have confined our attention to the variables that are considered the most important for the marking task based on the experience of human analysts, as indicated in the third column. In particular, we have not considered the pulse shape (as given by the pulse width at 3 dB, 6 dB, and 12 dB) or the estimated noise level, which could potentially lead to improvements in the marking accuracy. In future work, it would be of interest to develop models that can leverage this additional information.

```
*Peak,Hyfon,Chan,PW,PW,PW,Peak,Noise,Auto,Track,TagID,Tag
*Loc.,No.,No.,3dB,6dB,12dB,Amp.,Level,Thresh,Type,No.,Type
* Start Sequence at Sun Nov 17 00:00:00 2019 Peak Location = 0
69435,1,1,16,24,36,2124,2124,2124,USER,9152.00-24,VSTND
76974,1,1,11,26,39,1491,1491,1491,USER,9152.00-24,VSTND
93288,1,1,11,20,30,21726,21726,21726,USER,29401.00-04,VSYNC
96779,1,1,11,20,30,21634,21634,21634,POST,29401.00-04,VSYNC
96895,1,1,27,39,55,2492,2492,2492,USER,29401.00-04,VSYNC
179262,1,1,15,24,35,1946,1946,1946,USER,9152.00-24,VSTND
186800,1,1,12,21,32,1163,1163,1163,USER,9152.00-24,VSTND
289088,1,1,14,24,35,1938,1938,1938,USER,9152.00-24,VSTND
296628,1,1,17,27,49,1824,1824,1824,USER,9152.00-24,VSTND
324874,1,1,13,21,32,430,430,430,USER,29201.00-02,VSYNC
327850,1,1,14,21,28,362,362,362,POST,29201.00-02,VSYNC
398914,1,1,16,28,41,1933,1933,1933,USER,9152.00-24,VSTND
406454,1,1,14,26,38,1491,1491,1491,USER,9152.00-24,VSTND
437777,1,1,14,21,31,570,570,570,USER,30509.00-15,VSYNC
443246,1,1,21,34,48,529,529,529,POST,30509.00-15,VSYNC
446097,1,1,11,20,30,21622,21622,21622,POST,29401.00-04,VSYNC
```

Figure 2: A small excerpt of a marked acoustic data file.

Name	Description	Used in this work
Peak Location	Time stamp of the received ping determined with a 12,000 kHz clock. The Peak Location divided by 12,000 gives the arrival time of the ping measured in seconds since the hydrophone's most recent auto-reset. The hydrophone auto-resets every day at a specific time. This auto-reset time can be different for different hydrophones.	Yes
Hydrophone Number	The hydrophone identifier number.	Yes
Channel Number	The channel number of the hydrophone.	No
PW 3 dB, 6 dB, 12 dB	Pulse widths at -3 dB, -6 dB, and -12 dB	No
Peak Amp.	Peak amplitude of the pulse	Yes
Noise Level	Average noise level over a one-second period prior to receiving the ping	No
Auto Thres	Average adaptive threshold over a one-second period prior to receiving the ping	No
Track Type	Type of marking process. “POST” indicates the marking done using MarkTags software. “USER” indicates the manual marking done. “GPS SYNC” indicates the signals generated by the receiver to synchronize with UTC time.	Yes
TagID No. (Period No.)	The period of the fish tag. The format used is PPPP.SS where P indicates period and S indicates subcode.	Yes

Table 1: Information recorded for each received ping in the data files. Note that the last two variables are only presented in the marked data files (not in the raw data files).

MODEL DEVELOPMENT

Using the image representation discussed above, the problem of identifying the acoustic tags has effectively been transformed to the task of recognizing track-like features in an image, a time-consuming but accurately solvable task for a trained human operator. This is also a task that CNNs have proven highly successful at solving in recent years [Goodfellow et al., 2016]. In this project, we adopted a staged approach. In the first stage, we trained a CNN at determining whether a certain acoustic tag was present in a given image. This demonstrated that deep neural networks could be trained to recognize the tracks produced by the acoustic tag with high accuracy, but also revealed two failure modes. Data augmentation strategies were developed to mitigate these problems, as discussed in detail below. In the second stage, we trained a so-called U-Net [Ronneberger et al., 2015] at determining the precise location and extent of the tracks in the image. Here, the data augmentation methods developed in the first stage played an important role in enhancing the model performance.

Augmentation Methods

In this work, two augmentation methods were found to have a beneficial impact on the performance of the U-Net.

In the first augmentation technique, we add random offsets between 0 and the ping period to the time stamps, thereby generating multiple versions of the same image which only differ by a vertical translation. We found that this led to a significant improvement in the U-Net's ability to segment tracks irrespective of their vertical placement in the image.

The second augmentation technique was developed to help the U-Net to only identify the tag of interest and not get confused by tags with periods close to that of the tag of interest. In designing this augmentation technique, it is important to consider the effect of fish movement: a fish moving towards or away from the receiver generates pings separated by a time interval that is slightly shorter or longer than the actual period, respectively. This effect is also referred to as the Doppler effect, and, for movement at constant velocity, results in a ping track with non-zero slope. Fish rarely swim at a constant velocity, however, but change speed and direction often so the observed tracks can be rather complex and only approximate a straight line when viewed at coarse resolution.

For example, a fish with a tag period of 9 s moving towards the receiver at a speed of 1 m/s will have its apparent period reduced by approximately 1 / 1,500 x 9 s = 6 ms, where the speed of sound in water is 1,500 m/s. Therefore, assuming 1 m/s as an upper limit on attainable fish speeds, we may conclude that observed periods should be within 6 ms of the expected period. Based on this estimate, we augment the images as follows: for each image, we generate two related images by computing the remainder with respect to periods that are 10 ms shorter and 10 ms longer than the actual period, and label both of these images as "negative." We note that the chosen shift of 10 ms depends on the tag period and the assumed maximum swim speed of the tagged fish. In particular, faster fish may necessitate larger shifts.

We note that the above augmentation techniques were only applied to the training data. The test data were analyzed without alteration.

U-Net

In digital image processing, image segmentation is the process of dividing an image into segments, for example, identifying distinct objects like a cat or car in a real-world photo. Thus, segmenting an image means assigning labels to each pixel.

Here, we train a deep neural network architecture known as U-Net [Ronneberger et al., 2015] at identifying the near-horizontal tracks produced by the acoustic tags. The U-Net is a state-of-the-art deep neural network architecture for image segmentation which has been successfully applied to diverse data domains including medical imaging [Li et al., 2018], self-driving vehicles [Tran and Le, 2019], and satellite photography [McGlinchy et al., 2019]. The work presented here presents a novel adaptation of the U-Net to acoustic fish tracking data.

The U-Net assigns a score between 0 and 1 to each pixel in the image, where a high score indicates that the pixel likely forms part of a ping track while a low score suggests the opposite. For the purpose of obtaining a definite label, a fixed threshold value is commonly adopted to convert the score mask into a binary (yes/no) mask. A larger threshold value will generally result in fewer false identifications ("false positives"), but also result in more pings being missed ("false negatives") whereas a lower threshold value will have the opposite effect, increasing the number of false identifications but lowering the number of missed pings. In this work, a threshold value of 0.2-0.3 proved optimal.

Inverse Mapping: Pixels to Pings

As mentioned above, the U-Net outputs a mask of scores between 0 and 1, which may be converted to a binary mask using a constant threshold value.

When a pixel only contains a single ping, the reverse mapping from pixel to ping is straightforward. However, in some cases a single pixel contains multiple pings with near identical timestamps, creating ambiguity as to which ping should be attributed to the acoustic tag. Since the vertical pixel size is about 50 ms, this ambiguity can, in the very worst case, lead to substantial errors in positioning, theoretically up to 1,500 m/s x 50 ms = 75 m.

Such ambiguity may occur, e.g., due random noise signals or, more commonly, due to the detection of a multi-path signal, e.g., a reflection of the water surface or the seabed. Depending on the seabed depth and the relative positioning of the hydrophone and the acoustic tag, the delay between the direct-path signal and the multi-path reflections may be as short as a few milliseconds.

In order to filter out such noise signals and multi-path signals, we have implemented a simple rule-based algorithm that draws a trendline through the pings identified by the U-Net and discards any pings that deviate by more than a set tolerance from the trendline. In building the trendline, the algorithm initially considers only "high-fidelity" detection events in which both the primary and the secondary ping were detected with a temporal spacing

consistent with the known subcode. At every step, we update the vertical position and slope of the trendline by computing a weighted sum of the predicted position and slope (assuming a constant slope) and the observed position and slope. The relative weighting is determined as exp(-dt/T) where dt is the size of the time step and T is a fixed time constant, which effectively controls the "inertia" of the trendline. In a second pass, "low-fidelity" pings are recovered if they fall within a set tolerance of the trendline. If ambiguity persists, the earliest detected ping is used.

Model Training and Testing

To facilitate model development, we split the data into non-overlapping, six-hour segments, and further group these segments through random selection into three categories according to their use as training, validation or test data, in proportions of 60%, 20%, and 20%. We used six-hour long segments to reduce proximity between training and test samples and, thereby, train a robust and generalizable model. When generating images, we use a step size of 10 minutes, which implies a 20-min overlap between successive images. Thus, 34 different images are generated from each six-hour segment. The use of overlapping images within a six-hour segment allows us to make better use of the available data.

In the testing phase, the overlap between consecutive images implies that each pixel gets assigned (up to) three different scores by the U-Net. Before converting the scores to a binary mask, we average the available scores.

We trained a U-Net model with data of four different tags and six different hydrophones. The tags chosen are 9289.00-25, 9152.00-24, 9262.00-23, and 9316.00-25; and the hydrophones 1 to 6 are chosen for training. Among the four tags, the first three tags were chosen because they have the greatest number of pings compared to all the other tags in the study while the fourth tag was chosen because it had comparatively fewer pings. These choices were made to i) maximize the number of images containing marked pings, while ii) still exposing the U-Net to both frequently appearing and rarely appearing tags to achieve satisfactory performance across a wide range of tags. In order to obtain a reliable measure of the U-Net's ability to handle unseen data, we tested its performance on tags not used in the training phase.

Tables 2 and 3 provide a summary of the hydrophone-tag combinations used for training and validation, and testing, respectively. For training and validation, we combine the data from the various hydrophone-tag combinations into a single dataset, while for testing we keep the data separated. Note that some pings were not present near every hydrophone and the prevalence of a given tag varies greatly by hydrophone and temporally.

The U-Net was trained for 20 epochs (i.e., until it had seen every sample 20 times) although the accuracy on the validation set typically saturated already after the first few epochs of training. Standard values were used for U-Net hyperparameters throughout the training.

Computing Infrastructure

The model training and testing were done on the DeepSense high performance computing cluster with each training cycle or test using

Hydrophone No.	Tag period (ms)	Tag subcode (ms)	Number of marked pings	Tag presence (%)
1	9289	647	312,237	65
1	9152	628	338,387	69
1	9262	610	57,233	12
1	9316	647	0	0
2	9289	647	258,936	54
2	9152	628	191,644	39
2	9262	610	81,947	17
2	9316	647	0	0
3	9289	647	85,537	19
3	9152	628	98,403	20
3	9262	610	128,107	27
3	9316	647	7	0.000015
4	9289	647	107,182	22
4	9152	628	84,324	17
4	9262	610	82,584	17
4	9316	647	0	0
5	9289	647	9,474	2.0
5	9152	628	12,989	2.7
5	9262	610	46,569	9.7
5	9316	647	14,675	0.031
6	9289	647	12,675	2.6
6	9152	628	14,036	2.9
6	9262	610	39,672	8.2
6	9316	647	8,805	0.018

Table 2: Summary of the training set. For each hydrophone-tag combination, we list the number of pings marked by the human analyst used for training and the "tag presence" calculated as the ratio of marked pings to the total number of pings emitted by the tag within the time periods considered for training.

Name	Hydrophone No.	Tag period (ms)	Tag subcode (ms)	Number of marked pings	Tag presence (%)
Test-1	1	9234	710	7061	0.88
Test-2	2	9891	688	21708	2.9
Test-3	3	9508	548	31446	4.0
Test-4	4	9234	710	11298	1.4
Test-5	5	9180	733	31052	3.8
Test-6	6	9508	548	27765	3.6

Table 3: Summary of the test sets.

a 20 Core IBM Power8NVL 4.0 GHz compute node with 512 GB of RAM and a pair of NVIDIA Tesla P100 GPUs with 16 GB of GPU memory. It took about an hour and thirty minutes to train the U-Net on the DeepSense high performance computing cluster compared to nearly 20 hours on a laptop, which was done initially.

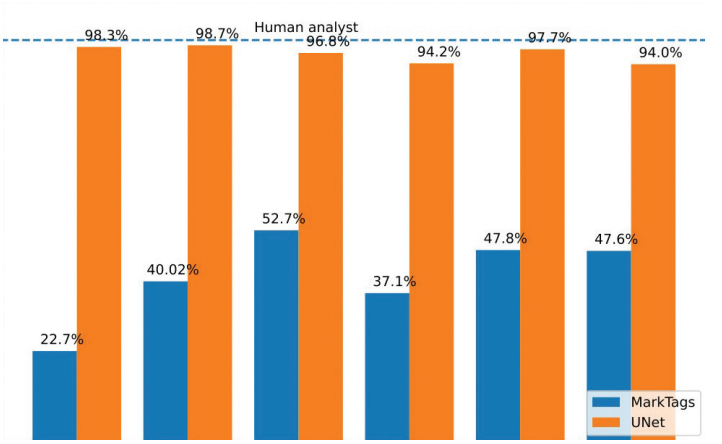


Figure 3: Marking accuracy of the U-Net solution proposed in this work compared to the existing auto-marking solution of the MarkTags software.

Performance Metrics

To quantify the model's performance, we use the F1-score [Goutte and Gaussier, 2005], defined as the harmonic mean of precision and recall, i.e., F1 = 2P*R/(P + R), where R is the recall, i.e., the fraction of the pings of interest that were marked, and P is the precision, i.e., the fraction of the marked pings that were in fact pings of interest. Thus, the F1 score considers both recall and precision, attaching equal importance to the two.

RESULTS

In Figure 3, we compare the performance of the U-Net on the six test sets to the MarkTags auto-marking software. In all the cases, the U-Net model outperforms MarkTags by a large margin, deviating only by a few percent from the human analyst. The U-Net achieves an accuracy of (96.6 +- 1.9)%, while the MarkTags auto-marking solution only achieves an accuracy of (41 +- 10)%.

It is worth noting that the manual marking effort involves a certain level of subjectivity, i.e., there are cases in which the identification of pings is ambiguous and the human analyst has to make an "educated guess" at which pings to mark. Indeed, in testing the performance of the U-Net we came

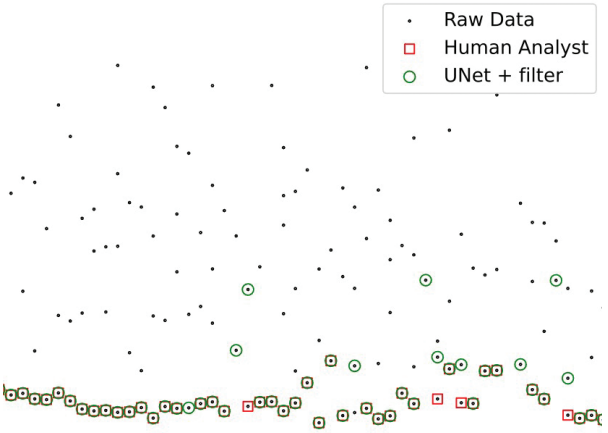


Figure 4: Zoom-in of the region enclosed by the box in Figure 1, with the pings marked by the human analyst indicated by empty squares (red), and the pings marked by the U-Net after applying the inverse pixel->ping mapping indicated by empty circles (green).

across cases in which there was substantial disagreement between the U-Net and the human analyst, but the human analyst appeared to have made an incorrect assessment. An experienced analyst was tasked with performing a blind analysis of these cases and when their effort was compared to the U-Net's prediction, it was found that the results were considerably closer to the U-Net's predictions, with the F1-score increasing from 74% to 85% on these remarked intervals.

An example of the image representation used to identify the acoustic tags was shown in Figure 1. In Figure 4, we show a zoom-in of the region enclosed by the box, with the pings marked by the human analyst indicated by empty squares (red), and the pings marked by the U-Net after applying the inverse pixel->ping mapping indicated by empty circles (green). Overall, we observed a high level of agreement between the two, although in a few cases there is disagreement. In particular, there are three high-lying pings that have been marked by the U-Net although they are located somewhat above the trend line. These pings likely represent multi-path signals and, therefore, were not marked by the human analyst. By adjusting the parameters of the inverse-mapping filter, in particular the "deviation tolerance," we can ensure that these pings are not marked by the U-Net.

Faint Tracks

Upon closer examination of the test sets, we have found that the U-Net performance is uneven across the test sets. In particular, the U-Net was unable to detect very faint tracks and also struggled to determine the precise start and end points of well-resolved tracks. An example of a faint track, in which only a small fraction of the transmitted pings are detected by the receiver, is shown in Figure 5. The image also contains a substantial number of pings from other acoustic tags, which makes this a particularly challenging case. As one can see, the U-Net was only able to detect a subset of the pings marked by the human analyst.

Detecting faint tracks and determining the precise start and end points of tracks is particularly important for fish movement reconstruction, which requires the acoustic tag to be detected independently by at least three hydrophones. By improving the U-Net's ability to detect faint tracks, acoustic tags will become detectable at farther ranges, allowing us to cover a given study area with fewer

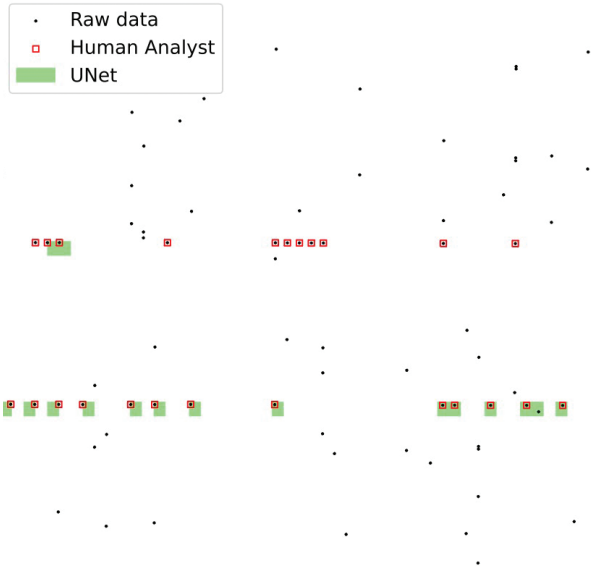


Figure 5: An example of a set of faint tracks which the U-Net struggles to identify.

hydrophones, thereby, reducing deployment costs significantly. For this reason, our future efforts will focus on improving the U-Net's ability to detect faint tracks.

CONCLUSION AND OUTLOOK

Acoustic tracking technology can be used to gain a better understanding of fish movement, ultimately allowing fisheries to optimize both operations and sustainability. Studies that require large numbers of tags in one place at one time are increasing in numbers and importance as we study smaller fish and their interactions with human modifications to their habitats [Muñoz et al., 2020; Aspillaga et al., 2021]. However, current tracking solutions are susceptible to noise and interfering signals and, hence, do not work well for such studies. A novel solution in which tags are identified solely from their ping rate has the potential to overcome these challenges, but the data generated by this solution require manual verification, which has made it difficult to operationalize the technology.

In this work, we have shown deep learning methods can help automate the analysis of the acoustic data. Leveraging a large, manually annotated, acoustic dataset and employing several data augmentation techniques, we have trained a neural network architecture known as U-Net to detect acoustic fish tags. The dataset derives from a month-long study performed in a shallow river environment, employing several dozen tags and 10 receivers. When the trained U-Net was asked to identify tags not seen during the training phase, it was able to do so with an accuracy of over 95%, which is close to the accuracy achieved by human analysts.

Employing modest computational resources, we were able to process the full dataset in a matter of hours, whereas it takes about 24 hours for one person to manually mark the data collected in just one day. Thus, the automated solution provides a considerable advantage in terms of both pace and effort compared to manual marking.

Several steps were taken to ensure a robust and generalizable model. This included splitting the data into six-hour segments to reduce proximity between training and test samples, application of several data augmentation strategies, and testing the model on unseen tags. Therefore, we believe that our work provides a good framework for training deep learning models to recognize patterns in acoustic data.

Our approach requires prior knowledge of the tag period to create images. Developing a model that does not require prior knowledge of the tag period would be relevant for some applications such as detecting tagged fish that have travelled long distances. A possible solution would involve detecting acoustic tags of unknown periods by computing the autocorrelation of the raw ping time series, although such an approach is unlikely to succeed if the signal to noise ratio is low or the tag experiences frequent and large Doppler shifts. A double-pass solution that combines the autocorrelation with the U-Net could potentially alleviate some of these difficulties.

Another, mostly technical, limitation of the current approach is that the U-Net only marks a single acoustic tag at a time (in a single hydrophone). A solution that could mark multiple acoustic tags simultaneously would be of interest from the point of view of ease of use and possibly faster computation.

Finally, we have found that the U-Net struggles at detecting faint tracks as well as determining the precise end points of well-resolved tracks. Resolving this issue is particularly important for fish movement reconstruction and will be pursued next.

It remains to be established how well the U-Net trained on this particular dataset generalizes to entirely new acoustic environments and study settings, for example, noisier environments in which multi-path signals occur more frequently. This includes environments where turbines are present, ocean environments affected by strong tidal currents, or river environments with rapids. Thus, future work would involve systematic investigation of generalization capacity to other datasets.

While the U-Net is a highly appropriate architecture for solving the acoustic tag identification problem, it may not necessarily be the optimal choice. Therefore, future studies should also explore the efficacy of alternative networks. However, the results obtained so far are already highly encouraging. Our work demonstrates that deep learning can, if not entirely eliminate, drastically reduce the need for manual analysis of the data generated by the new encoding scheme, thus paving the way for large-scale application to fisheries operation. We are now working on improving and optimizing the U-Net model so that it can be integrated into daily operations.

ACKNOWLEDGMENTS

This work was funded by Mitacs through the Accelerate program in partnership with Innovasea. This research was enabled in part by support provided by Innovasea Inc. (www.innovasea.com), and computations were performed on the DeepSense (www.deepsense.ca) high-performance computing

platform. We thank Jennifer LaPlante and Lu Yang of DeepSense for their support in project administration and computing resources.

REFERENCES

Aspillaga, E.; Arlinghaus, R.; Martorell-Barceló, M.; Follana-Bernám G.; Lana, A.; Campos-Candela, A.; and Alós, J. [2021]. *Performance of a novel system for high-resolution tracking of marine fish societies*. Animal Biotelemetry, Vol. 9, No. 1.

Ehrenberg, J.E. and Steig, T.W. [2003]. *Improved techniques for studying the temporal and spatial behaviour of a fish in a fixed location*. ICES Journal of Marine Science, Vol. 60, pp. 700-706.

Ehrenberg, J.E. and Steig, T.W. [2009]. *A study of the relationship between tag-signal characteristics and achievable performances in acoustic fish-tag studies*. ICES Journal of Marine Science, Vol. 66, pp. 1278-1283.

Goodfellow, I.; Bengio, Y.; and Courville, A. [2016]. *Deep learning*. MIT Press. www.deeplearningbook.org.

Goutte, C. and Gaussier, E. [2005]. *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation*. Advances in Information Retrieval, Springer Berlin Heidelberg, pp. 345-359.

He, K.; Zhang, X.; Ren, S.; and Sun, J. [2015]. *Delving deep into rectifiers: surpassing human-level performance on ImageNet classification*. IEEE International Conference on Computer Vision (ICCV), pp. 1026-1034.

Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; and Kingsbury, B. [2012]. *Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups*. IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97.

Hussey, N.E. Kessel, S.T.; Aarestrup, K.; Cooke, S.J.; Cowley, P.; Fisk, A.; Harcourt, R.G.; Holland, K.N.; Iverson, S.J.; Kocik, J.F.; Mills Flemming, J.E.; Whoriskey, F. [2015]. *Aquatic animal telemetry: a panoramic window into the underwater world*. Science, Vol. 348, 1255642.

Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. [2018]. *H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes*. IEEE Transactions on Medical Imaging, Vol. 37, No. 12, pp. 2663-2674.

McGlinchy, J.; Johnson, B.; Muller, B.; Joseph, M.; and Diaz, J. [2019]. *Application of UNet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery*. IGARSS 2019 – IEEE International Geoscience and Remote Sensing Symposium, pp. 3915-3918.

Muñoz, L. Aspillaga, E.; Palmer, M.; Saraiva, J.; and Arechavala-López, P. [2020]. *Acoustic telemetry: a tool to monitor fish swimming behavior in sea-cage aquaculture*. Frontiers in Marine Science, Vol. 7, p. 645.

Ransom, B.; Steig, T.; Timko, M.; and Nealson, P. [2008]. *Basin-wide monitoring of salmon smolts at US dams*. International Journal on Hydropower and Dams, Vol. 15, pp. 44-49.

Ronneberger, O.; Fischer, P.; and Brox, T. [2015]. *U-Net: convolutional networks for*

biomedical image segmentation. In: Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. (eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351. Springer, Cham.

Semmens, B.X. [2008]. *Acoustically derived fine-scale behaviors of juvenile Chinook salmon (Oncorhynchus tshawytscha) associated with intertidal benthic habitats in an estuary*. Canadian Journal of Fisheries and Aquatic Sciences, Vol. 65, pp. 2053-2062.

Steig, T.W.; Skalski, J.R.; and Ransom, B.H. [2004]. *Comparison of acoustic and PIT tagged juvenile chinook, steelhead and sockeye salmon (Oncorhynchus spp.) passing dams on the Columbia River, USA*. In: Aquatic telemetry-advances and applications, Proceedings of the Fifth Conference on Fish Telemetry in Europe, Ustica, Italy, 9-13 June 2003, pp. 275-286.

Tran, L-A. and Le, M-H. [2019]. *Robust u-net-based road lane markings detection for autonomous driving*. Proceedings of 2019 International Conference on System Science and Engineering, ICSSE, pp. 62-66.