Version of Record: https://www.sciencedirect.com/science/article/pii/S0378383921000193
Manuscript_8a5ed6593891c7ab5e292b7ae07c8142

Automated Rip Current Detection with Region based Convolutional Neural Networks

Akila de Silva[a], Issei Mori[a], Gregory Dusek[b], James Davis[a] and Alex Pang[a]

[a]*University of California, Santa Cruz, CA, United States*
[b]*NOAA National Ocean Service, Silver Spring, MD, United States*

ARTICLE INFO

Keywords:
Rip Current Detection
Machine Learning
Image Processing
Faster R-CNN
Temporal Smoothing

ABSTRACT

This paper presents a machine learning approach for the automatic identification of rip currents with breaking waves. Rip currents are dangerous fast moving currents of water that result in many deaths by sweeping people out to sea. Most people do not know how to recognize rip currents in order to avoid them. Furthermore, efforts to forecast rip currents are hindered by lack of observations to help train and validate hazard models. The presence of web cams and smart phones have made video and still imagery of the coast ubiquitous and provide a potential source of rip current observations. These same devices could aid public awareness of the presence of rip currents. What is lacking is a method to detect the presence or absence of rip currents from coastal imagery. This paper provides expert labeled training and test data for rip currents. We use Faster R-CNN and a custom temporal aggregation stage to make detections from still images or videos with higher measured accuracy than both humans and other methods of rip current detection previously reported in the literature.

1. Introduction

Rip currents are the most significant safety risk to swimmers along the coastlines of oceans, seas, and large lakes. [9, 10, 16]. The majority of beach goers do not know how to identify rip currents, and there is no robust and reliable location-independent method to identify them. Globally there are thousands of drownings each year due to rip currents [28, 56]. A 20 year study by the US Lifesaving Association reports that 81.9% of the 37,000 beach rescues each year are due to rip currents [9]. Lifeguards are often trained to identify rip currents. However the majority of drownings occur on beaches without trained personnel [2, 6]. Posted signs can provide a warning, but there is evidence that most people do not find existing signs helpful in actually identifying rip currents [8]. There has been no decline in the number of associated drowning fatalities, despite warning signs and educational material.

Rip currents are a well-studied ocean phenomenon [5, 36, 38]. They are defined as strong and narrow channels of fast-moving water that flow towards the sea from beaches. When waves break, they form a "setup" or an increase in mean water level. This setup can vary along a shoreline depending on the amount or height of breaking waves. Rip currents form as water tends to flow alongshore from regions of high setup (larger waves) to regions of lower setup (smaller waves) where currents converge to form a seaward flowing rip. Furthermore, macrovortices, induced by alongshore uneven wave breaking, may also be a contributing factor for rip current formation and evolution. [11, 46, 68, 66]. The speed of seaward rips can be quite strong reaching 2 m/s, faster than an Olympic swimmer. There are multiple factors that determine the location and strength of rips, such as bathymetry, wave height and direction, tide, and beach shape. Rip currents may either be transient or persistent in space and time. Rips that are frequently found at the same location are usually indicative of a fairly stable bathymetric feature such as a sand bar or reef, or a hard structure such as rocky outcrop, jetty or pier. These bathymetric features results in variations in wave breaking and setup leading to channelized rip current flow. Transient or flash rips are independent of bathymetry and may move up or down the beach, and may appear or disappear. Transient rips are best understood with respect to vorticity due to short-crested wave breaking and the subsequent eddy coalescence [48, 49].

The Southeast Coastal Ocean Observing Regional Association (SECOORA) in partnership with the National Oceanic and Atmospheric Administration (NOAA) maintains a network of coastal web cameras for different applications such as montoring wave runup, human use of natural resources, and spotting rip currents [24]. These data are supporting the validation of a rip current forecast model to alert people to potential hazards [25]. The most commonly

ORCID(S):

Figure 1: A collection of beach scenes, some of which contain rip currents. Unfortunately these rip current "objects" do not have clear shape, and most people find them hard to identify.

used method to visualize rip currents from video is time averaging, summarizing a video as a single image [35]. However these time averages when manually assessed can be misinterpreted. Furthermore, they are not readily available nor interpretable by the average beachgoers, and the process of averaging removes available information.

In recent years the coastal engineering community has successfully used deep neural networks to solve many problems. Classification problems such as classifying wave breaking in infrared imagery [12], beach scene and other landscape classification [14], automated plankton image classification [55] and ocean front recognition [50] were formulated as deep learning problems using convolutional neural networks. Furthermore, some regression problems such as optical wave gauging [13], tracking remotely sensed waves [74], typhoon forecasting [42] were also solved using deep neural networks. In addition, generative adversarial networks, a type of deep neural networks, were used to improve the quality of downscaling of ocean remote sensing data [23].

Object detection with deep neural networks is well studied in the computer vision community. However most benchmarks and research focus on detecting physical objects with boundaries between what is and is not an object [20, 27, 51]. Rip currents are ephemeral "objects" which are not observable in every frame, and amorphous without clearly defined boundaries even when observable. It is not clear whether existing methods are applicable. Figure 1 provides a set of examples, illustrating the difficulty of the problem. In some of the images rip currents are clearly visible while in some it is difficult for a layperson to recognize the presence of a rip current.

Our work is aimed at introducing this problem to the coastal engineering community, and showing that object detection methods *are* applicable. We gathered training data of rip currents and labelled those with bounding boxes indicating the location of the rip current with a co-author who is a rip current researcher at NOAA. We use Faster R-CNN [71] with a custom temporal aggregation stage that allowed us to achieve detection accuracy that is higher than both humans and other methods of rip current detection previously reported in the literature.

The remainder of the paper is organized as follows. All the related work is summarized in Section 2. We discuss how the data was collected in Section 3. Our method is discussed in 4. Results are discussed in 5. Limitations and discussion are in Section 6. In Section 7 we conclude our paper. And in Appendix A we provide the link to the supplementary materials.

2. Related Work

Rip currents can be observed using both in situ and remote sensing methods. Among the in situ methods, wave sensors, acoustic velocimeters, or current profilers can be deployed at specific locations [26, 40, 43, 54]. Floating drifters with embedded GPS units have also been used to measure currents [15, 16, 73]. These methods are costly, time consuming, require technical expertise and are generally only applicable to highly localized instances in time and space. These limitations severely hinder the applicability of such approaches to both public warnings and model validation. In comparison, remote sensing technologies such as satellite, air-borne, and ground-based imaging as well

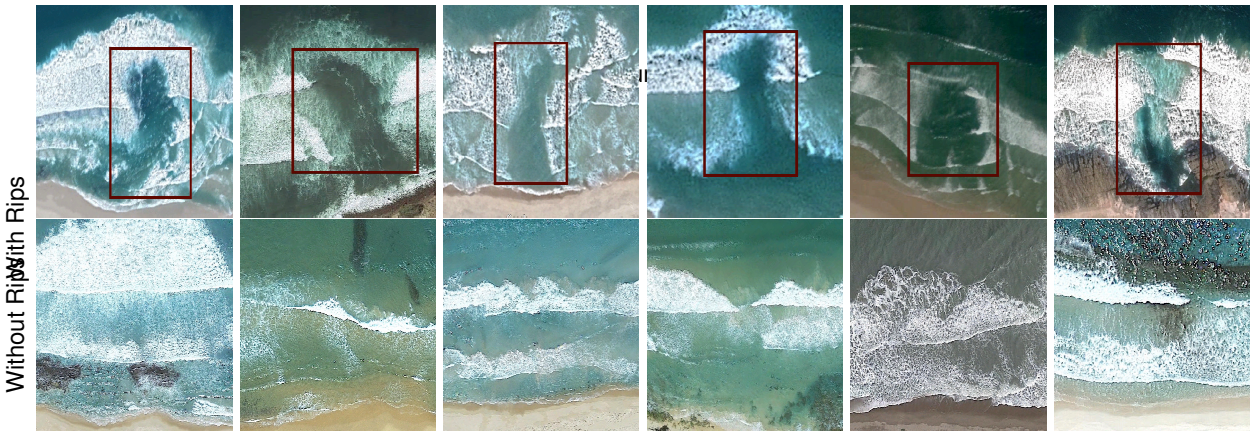

Figure 2: Examples drawn from the 2440 images we collected and labeled to build a training data set. Ground truth bounding boxes are shown in red.

as radar imaging provide more coverage and less expensive (e.g. web cameras) alternatives [31, 60]. A hybrid approach of adding fluorescein dye into the water and observing its dispersion using aerial video provides dramatic visualization of rip currents [7, 18, 19, 69].

Time averaged images are a routine method for analyzing video in oceanic research, with 10 minutes being a common integration period [34, 35, 52, 67]. This method is popular because averages often make identification of rip channels easier for the human eye. While these images are usually intended for human interpretation, Maryan et. al. apply shallow machine learning to recognize rip channels in time averaged images [59]. Nelko also used time averaged images and noted that prediction schemes developed at one beach location may not be directly applicable to another without some modifications [61]. Haller introduced wave averaging to enhance the detection of rip currents in microwave data; an approach which could also be applied to imagery data [31].

Dense optical flow [4, 37] has been used to detect rip currents in video [65]. This method is attractive since optical flow fields can be directly compared against ground truth flow fields obtained from in situ measurements [21]. Unfortunately these methods are sensitive to camera perturbation, and have difficulty in areas lacking textural information.

Certain kinds of rip currents are characterized by visible sediment plumes. These can be segmented based on hanges in coloration. For example, Liu et al., use thresholding in HSV color space to detect rip currents [54]. Unfortunately, not all rip currents contain sediment plumes, and thus this method is not applicable to our data sets.

Object detection in images is well studied in the computer vision literature [32, 53, 63, 70, 76]. These methods have been extended to detect objects in videos [33, 45, 79]. However, with the exception of Maryan et al. [59] which performs detection based on time averaged images, and Liu et al. [54] which performs detection based on color segmentation, they have not been applied to crisp images or videos for the purpose of rip current detection.

3. Data sets

3.1. Training Data

Since rip currents are a new problem domain for computer vision, we did not find any existing public databases of rip current images. Therefore, we assembled a training data set of rip current images and non-rip current images from scratch. Our primary source for the database was Google Earth^{TM}, which allowed us to extract high-resolution aerial images of rip currents and non-rip currents. In total, the database contains 1740 images of rip currents and 700 images of similar beach scenes without rip currents. The images range in size from 1086×916 pixels to 234×234 pixels. We annotated ground truth in the rip current images with axis-aligned bounding boxes: where the x axis and y axis of the bounding box is aligned with the x axis and y axis of the image. Some examples of the training data set are shown in Figure 2. Note that this data set contains unambiguous easy examples. We used this training data for training models described in Section 4.

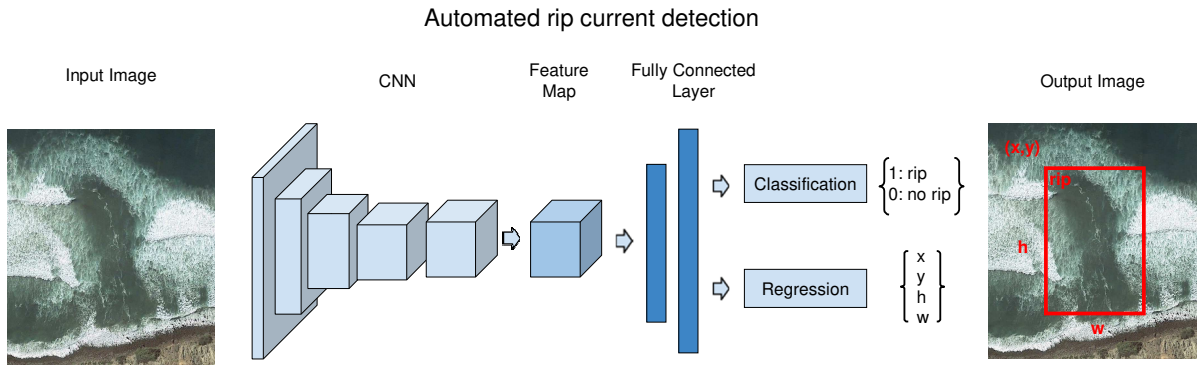

Figure 3: Architecture of Faster R-CNN: Classification and Regression branch. CNN represents the convolutional neural network.

3.2. Test Data

We also collected a data set of 23 video clips consisting of 18042 frames in total. Of those, there are a total of 9053 frames with and 8989 frames without rip currents. Image size varies from 1280 × 720 pixels to 1080 × 920 pixels. We used the bounding boxes labelled on each image as the ground truth. Figure 1 contains both positive and negative example frames, as well as a few frames from the training set that might be mistaken as containing a rip current. Note that this set contains more difficult cases.

The frames of this video data set were used for testing. Note that the static images in the training set were taken from high elevation while the videos used to test the model were taken from a lower perspective. Even so, the trained model performed well on the test frames from the video collection.

4. Method

4.1. Deep Learning

Deep learning methods have recently been used in many computer vision tasks. These methods have taken over many traditional shallow machine learning methods such as support vector machines, random forests, etc. [1, 41, 44, 77, 78, 82]. However, deep learning models require careful tuning of hyperparameters, which determine the deep learning model architecture. Multiple sensitivity analysis experiments are often needed to optimize hyperparameters for a particular model and task. Also, deep learning models are costly to train, usually taking days and many dedicated computation resources such as GPUs (Graphis Processing Units) or TPUs (Tensor Processing Units).

Unlike deep learning methods, traditional machine learning methods require a fair amount of human feature engineering by a domain expert, such as the features created by Maryan et al. [56] for the rip current detection task. Features engineered by humans may not be the most optimal. In deep learning, the algorithm learns the most optimal features through gradient decent that are necessary for the given task.

Even with a high computation cost of training, deep learning models learn a large number of parameters compared to traditional machine learning models, resulting in a higher accuracy in complex vision problems. Deep learning methods such as region based convolutional neural networks have out-performed traditional machine learning methods in many vision tasks, such as object detection [30, 59, 64]

4.2. Static Image Detector: Faster R-CNN

Region-based convolutional neural networks have achieved great success in object detection problems. These object detection models usually consist of separate classification and localization networks with a shared feature extraction network. In the computer vision community, Faster R-CNN is generally considered as the most accurate object detector [22]. Many critical applications such as fire detectors [3] and pedestrian detectors [47], successfully use Faster R-CNN. Also, many medical applications such as detecting cervical spinal cord injury and disc degeneration detection [57], breast cancer detection [58], malaria cell detection [39] use Faster R-CNN as their underlying object detector. Therefore, we choose Faster R-CNN as our single image detector.

The first is the deep convolutional neural network that proposes regions. The second is the Fast R-CNN detector [29]. Faster R-CNN follows the traditional object detection pipeline. It first generates region proposals, and then

categorizes each proposal as either rip current or background. Secondly, the classified bounding boxes are further refined. Essentially, the model learns a mapping from the generated regions to the actual ground truth with a regression network. The model then uses this mapping "function" during testing to refine the generated region. These refined bounding boxes can be anywhere in a frame as features are translation invariant [71]. If there is more than one bounding box detected in a frame, we only keep track of the largest one and ignore any additional boxes.

As shown in Figure 3 the convolutional neural network (CNN) consists of five hierarchical blocks. Each block consists of convolutional layers followed by a max pooling layer. The convolutional layers generate features by applying filters to the input. After which the output is fed into a max pooling layer where the feature map is down-sampled by only keeping the maximum values within a region. These two steps are repeated in the five blocks of the CNN, resulting in the final feature map.

Multiple regions from the input image are then projected onto the feature map. These regions are generated as a rectangle with predefined proportions and positions. These regions have a predefined area of 64^2 pixels, 128^2 pixels and 256^2 pixels. Each region is generated as an rectangle with three different aspect ratios for length and width: $1\ 1$, $1\ 2$, $2\ 1$, resulting in nine different regions. These regions are then positioned on a regular grid by generating the nine regions centering round each grid point. The regions that fall beyond the boundary of the image are ignored. The associated features for these regions are then fed into the fully connected layers where the decisions are made. The classification branch predicts if the region is a rip current or not. If the classification branch classifies the region as a rip current then the regression branch further refines the position and the size of the detected rip current.

We trained the Faster R-CNN model with the training data discussed in section 3.1. Before training, each image was augmented by rotating 90^0 degrees clockwise and counter clockwise, producing a training data set three times the size of the original training data set. All the training data was resized to 300×300 pixels before training to save computation time. We used bi-cubic interpolation to resize the images.

4.3. Frame Aggregation

Static object detection models only consider the information in the frame currently being processed. However, rip currents are natural ocean phenomena, with shape and texture change depending on many external factors such as weather, wind speed, wave field characteristics, water flow speed, floating debris, and dirt sediments. The exact boundaries of a rip current are not well defined. This is different than objects with well-defined edges such as pedestrians or vehicles. Applying detection algorithms to objects with amorphous boundaries such as rip currents produces bounding boxes with variable sizes and locations in adjacent video frames. In Figure 4 we illustrate this variability by drawing all correctly detected bounding boxes from one video sequence onto a single frame.

This variability affects overall accuracy, and would not instill confidence in the results if these bounding boxes were presented to a user as a video overlay. Thus we investigate temporal smoothing and aggregation to improve the results.

We find the overlapping regions of the detected bounding boxes by using an accumulation buffer with the same size as the input frame and initialized as a zero matrix. We consider a temporal window of N frames to build the accumulation buffer. In the first $N-1$ frames, the accumulation buffer is incremented by 1 for each region within a detection bounding box. Starting with frame N, the area covered by a detection bounding box is incremented by 1 and capped at a maximum of N. Regions not covered by a detection bounding box are decremented by 1, but retains a minimum of 0 in the accumulation buffer. In effect, the accumulation buffer keeps track of the bounding boxes using a sliding window of N frames. The process of building the accumulation buffer is illustrated in Figure 5, where areas with higher values are displayed as darker regions in the accumulation buffer. For purposes of identifying a single bounding box over the collection of bounding boxes across N frames, we consider only the regions of the accumulation buffer where the value is at least T, and draw the tightest possible axis-aligned bounding box around this region (see Figure 6). This is the aggregated detection. In our implementation we use N=60 and T=30.

Before frame aggregation, large variations in bounding box size occur in almost all consecutive frames, shown in Figure 7 top. After frame aggregation, the average size change is much smaller, with most frames having zero change in size from the prior frame. The variation in position of the bounding boxes is similarly reduced by frame aggregation, as seen in Figure 7 bottom. This improved temporal coherence provides a smoother and more consistent portrayal of the rip current location when shown as an overlay on the video.

When analyzing video clips, our method analyzes individual frames and places a bounding box around detected rip currents. Since analysis is on a per frame basis, the bounding boxes may move from frame to frame. In addition, it's possible that a rip current may not be detected in a particular frame. However, with our proposed frame aggregation

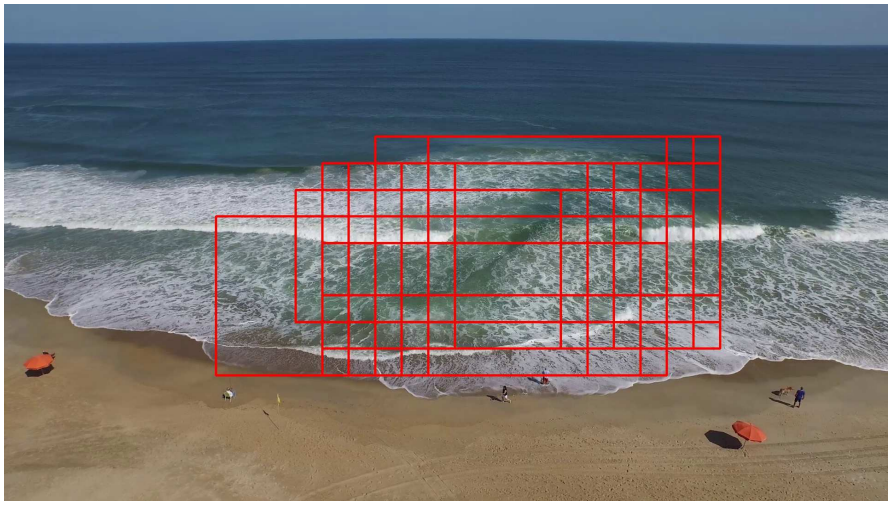

Figure 4: The bounding boxes marking the boundaries of detected rip currents from individual frames of a video segment are superimposed onto the most recent frame. Note how the bounding boxes from individual frames may move and/or change shape

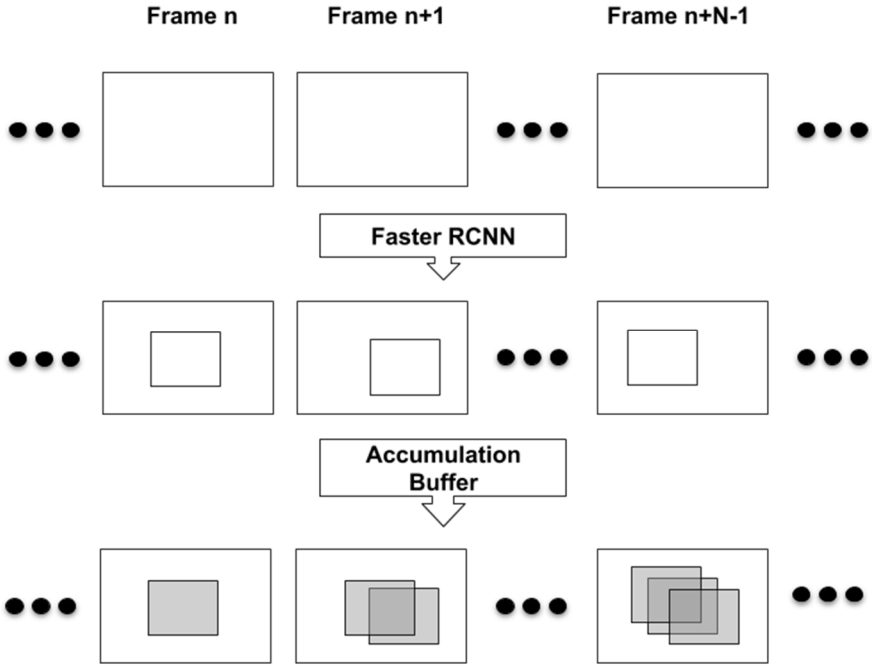

Figure 5: Frame aggregation for a window of length N. First row shows the input frame sequence. Second row shows the detections from Faster R-CNN. Third row shows the accumulation buffer.

strategy, instances when a few frames where the rip is not detected (false negative), or when there's a sudden change in the location or size of the bounding box (false positive) are both handled gracefully.

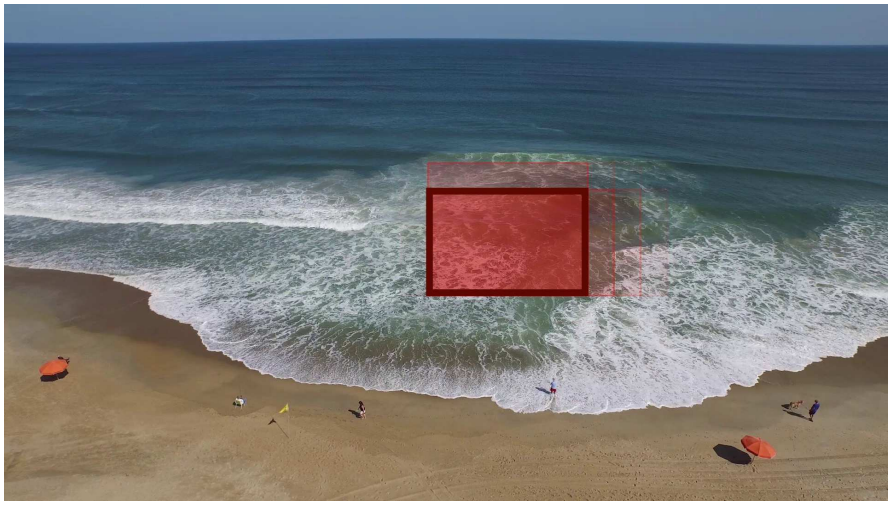

Figure 6: A visualization of the resulting accumulation buffer values. Regions with a higher value are shown in more opaque red. The resulting bounding box around thresholded values is shown in solid dark red. The full video can be seen in the supplementary material in appendix A

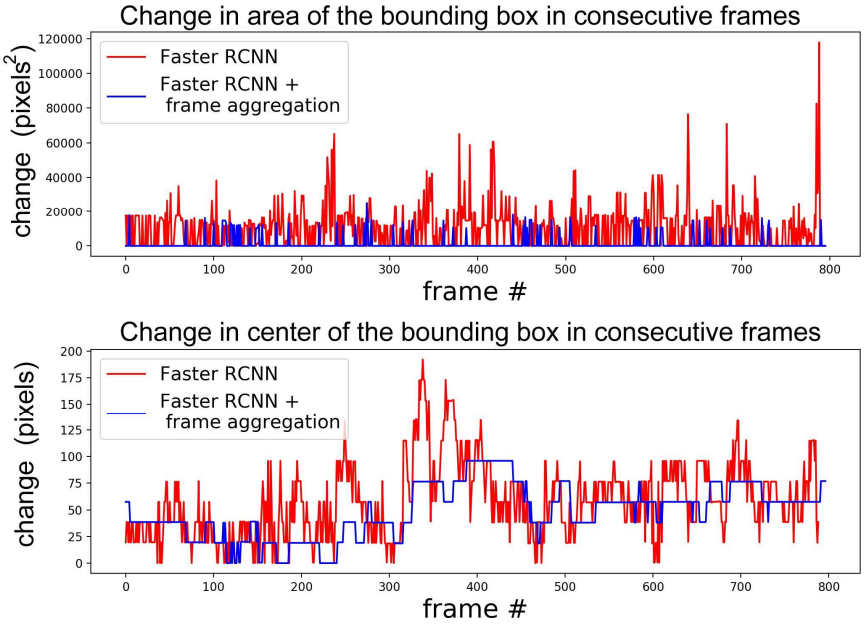

Figure 7: Plots showing the differences in area and center of bounding boxes in consecutive frames. Without frame aggregation (in red) there are much higher differences in bounding box sizes and positions than there is after frame aggregation (in blue).

5. Results

We compared the accuracy of our method with human observers as well as two prior methods. We also compared our own method with and without temporal aggregation.

Comparison metric. All methods were tested using the video data set. Frames were labeled as correctly classified if the detected bounding boxes have an Intersection over Union (IoU) [72] score versus ground truth above 0.3. IoU is

Figure 8: Rip current detections on some of the frames from the test data set. Red bounding boxes show the correctly detected rip currents. Blue bounding boxes show the ground truth. Frames without bounding boxes do not contain any rip currents.

calculated as *area_of_intersection/area_of_union* of the ground truth and the detected bounding boxes. Accuracy of the video was computed as *correct_labels/total_frames*, and Table 1 provides the results for all methods. Since rip currents do not have a well-defined boundary the scale of ground truth bounding boxes has some uncertainty. Therefore, the predicted bounding box does not need to closely match with the ground truth bounding box for a detection to be marked as correct. Because of this we choose a lower IoU threshold than detecting objects with well defined boundaries such as cars or tanks [71]. We visually verified that even with a lower IoU threshold the detected bounding box shows the location of the rip current. Also, we used this IoU threshold across all methods in Table 1.

Humans. One of the primary reasons for an automated method of rip current detection is that most people are not good at identifying rip currents [8]. To measure human accuracy of identifying rip currents, annotators were asked to draw bounding boxes around places they believe to have rip currents. We sampled every tenth frame from our video test set and randomized presentation order across all positive and negative examples. Human annotators were not carefully trained, instead they were provided three positive and three negative examples, roughly the amount of information which might fit on a sign at the beach. Annotators were acquired using Mechanical Turk, an online market where jobs are posted for workers [62, 17], with basic screening for reliable workers, and paid $0.10 per image.

Although human performance was relatively poor with only 76% of frames labeled correctly, it was higher than we expected based on past studies [8]. We hypothesize that the sample images showing both locations of rip currents and absence of rip currents are more effective than warning signs alone.

Time averaged images. Maryan et al. [59] perform detection on time averaged images using a boosted cascade of simple Haar like features [76]. We used Maryan's time averaged data for training since our training data set consists

.	Human	Philip [65]	Maryan [59]	Maryan [modified]	F-RCNN [ours]	F-RCNN+FA [ours]
*rip*_01.*mp*4	0.976	0.895	0.358	0.460	0.966	1.000
*rip*_02.*mp*4	0.700	0.098	0.698	0.135	0.776	0.860
*rip*_03.*mp*4	0.231	0.347	0.194	0.540	0.831	0.950
*rip*_04.*mp*4	0.757	0.800	0.487	0.780	0.939	0.970
*rip*_05.*mp*4	0.883	0.280	0.736	0.450	0.834	0.957
*rip*_06.*mp*4	0.881	0.000	0.167	0.470	0.753	0.890
*rip*_08.*mp*4	0.492	0.063	0.328	0.730	0.860	0.850
*rip*_11.*mp*4	0.824	0.000	0.563	0.940	0.930	0.951
*rip*_12.*mp*4	1.000	0.000	0.734	1.000	1.000	1.000
*rip*_15.*mp*4	0.967	0.137	0.315	0.390	0.760	0.870
*rip*_16.*mp*4	0.614	0.073	0.468	0.640	0.820	0.920
*rip*_17.*mp*4	1.000	0.321	0.064	0.750	0.980	1.000
*rip*_18.*mp*4	0.563	0.218	0.250	0.240	0.790	0.890
*rip*_21.*mp*4	0.901	0.543	0.486	0.180	0.940	1.000
*rip*_22.*mp*4	0.583	0.000	0.479	0.395	0.880	0.974
*no_rip*_01.*mp*4	0.986	0.169	0.000	0.972	0.813	1.000
*no_rip*_02.*mp*4	1.000	0.789	0.000	0.985	0.807	1.000
*no_rip*_03.*mp*4	0.919	0.000	0.000	0.981	0.984	1.000
*no_rip*_04.*mp*4	0.952	0.000	0.000	0.974	0.835	1.000
*no_rip*_05.*mp*4	0.903	0.000	0.000	0.986	0.833	1.000
*no_rip*_06.*mp*4	1.000	0.246	0.000	0.986	0.875	1.000
*no_rip*_07.*mp*4	0.983	0.525	0.000	0.982	0.875	1.000
*no_rip*_11.*mp*4	0.988	0.198	0.000	0.964	0.924	1.000
average accuracy	0.760	0.307	0.210	0.729	0.884	**0.984**

Table 1
Accuracy for each video in the test set. Column 3: Maryan[59] is trained on their training data. Column 4: Maryan [modified] is trained on our training data. The F-RCNN method has higher overall accuracy than humans or any of the prior methods tested. Frame aggregation does contribute to improvement in accuracy.

of only static frames. Testing was performed by first computing time averaged images on each video in our test data set. This method did not perform well. In order to determine if the cause was the images available for training or the model itself, we repeated the experiment with new data. We replaced the relatively small number of low resolution time-averaged images from [59] with the static images from our training data set. Testing this time was against single frames in our test data . This modification is called Maryan[modified] in Table 1. When using our training data the model accuracy improved considerably, leading us to conclude that appropriate training data is critical to good results. Furthermore, it suggests further investigation on whether using crisp images, rather than time averaged images, to train a model might produce more accurate results.

In our test images/videos the beach is always located at the bottom half of the image/video. However, the training data used by [59] was cropped from images where the beach is located in the top half of the image. Therefore, we were concerned that the difference of orientation between the training data and the test data contributed to the low accuracy of the model. However, when we retrained the model with vertically flipped training data we did not see any significant difference in accuracy on our test images/videos. We hypothesize that the reason for this is that the cropped training images contain insignificant amount of beach pixels.

Optical flow. Philip et al. [65] compute optical flow on video sequences and make the simplifying assumption that rip currents can be identified by regions with the second most predominant flow direction, after that of the primary incoming wave direction, and that they flow in a single seaward direction. This results in regions of actual rip currents, but also picks up swash regions where water is washed up the beach and back out to sea with the passing of each wave. This method was introduced with the primary intention of providing visualizations to users, rather than automated detection. To allow comparison, we modify the method to return a bounding box around the largest detected region, ignoring smaller regions which are less likely to be correct. This method performed poorly on our test data. We noticed

Figure 9: Example failure cases. The false positives on the beach scene (right) are not easily explainable. The false positive on the left scene happens only on spurious frames, which is then corrected by frame aggregation.

that in videos where there is not enough textural information on the rip current, the optical flow field generated was weak, leading to either missed detections or detection in other regions of the video with stronger texture.

Frame aggregation. We implemented frame aggregation as a post process to Faster R-CNN initially to temporally stabilize detections, driven by a need for user interpretable visualization of rip current location.

In order to understand whether temporal smoothing also increased accuracy we analyzed our implementation both with and without frame aggregation. We found that temporal aggregation leads to higher accuracy than using Faster R-CNN alone. Example detection results are shown in Figure 8. Numerical comparison of humans, prior methods, and our model are provided in Table 1. Faster R-CNN with frame aggregation had the highest accuracy in nearly all cases, and the highest overall (last column of Table 1). For visual comparisons we have added all the results in the supplementary materials at appendix A.

6. Discussion and Future Work

As with all machine learning models, our implementation can fail when used with images that do not resemble the training data set. Our data sets included primarily rip currents characterized by a gap in breaking waves, the most common visual indicator for bathymetry controlled rip currents. Thus we would expect to miss rip currents with other visual indicators like sediment plumes. We also expect to fail when presented with new imagery, and occasionally for no apparent reason at all, as seen in Figure 9.

Sensitivity Analysis. Sensitivity analysis experiments were conducted to determine the optimal values for parameters N and T used in frame aggregation (Section 4.3), training data and number of training iterations. Figure 10 shows the average accuracies of rip current detection from our test data using different values of N and T. The smallest $N(= 60)$ and $T(= 30)$ with the highest accuracy was chosen as the optimal N and T. By choosing the smallest N and T we are able to save on computation time as well.

Further experiments were conducted to determine the optimal amount of data needed to train the model. The model was trained with 25%, 50% and 100% of the training data. At each stage, the accuracy on the test data was recorded. Training the model with the full training data set lead to the highest accuracy as shown in Figure 11. Experiments were also conducted on when to stop training our model. The accuracy of the model on the test data was recorded at different training epochs. The training was stopped at epoch 60 when the accuracy plateaued as shown in Figure 12.

We did not conduct a sensitivity analysis of the hyperparameters of the network. We relied on the sensitivity analysis conducted by the authors of Faster R-CNN [71] to determine the network's optimal hyperparameters.

With frame aggregation the accuracy for test data without rip currents was high compared to the accuracy for test data with rip currents. Since frame aggregation takes into account the prior N detections, we can easily filter out spurious detections, leading to a higher accuracy for test data without rip currents.

We noticed that for one video, *rip*_08.*mp*4, frame aggregation did not improve the accuracy. For *rip*_08.*mp*4, the placement of bounding boxes identifying the rip current from individual frames by F-RCNN is more spread out. This

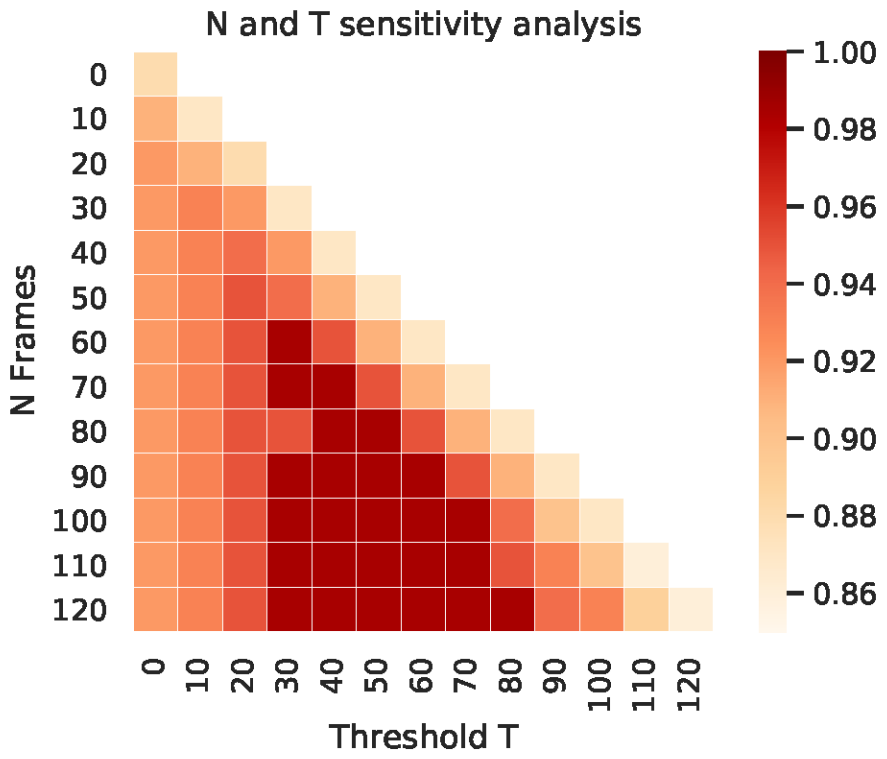

Figure 10: Heatmap of accuracy values for parameters N frames and threshold T.

leads to smaller over lapping regions of bounding boxes. When used with frame aggregation, the resulting bounding box is smaller than the bounding box without frame aggregation. This produces a more conservative estimate of the size of of the rip but similar "confidence" of its location using the same IOU threshold.

The frame aggregation method discussed in Section 4.3 works well when the rip current is fairly stationary relative to the camera. Specifically, with = 60 frames and a frame rate of 30 frames per second, we are assuming that the rip current is in the same location over a 2 second interval of the video. This assumption is generally true for video captured from stationary cameras such as CCTV or web cameras. If the video were captured using mobile devices such as smartphones or drones where the rip current moves a significant amount within the frame, additional information such as the camera's accelerometer and GPS can possibly be incorporated in the processing. This is a subject for future work.

We are aware that there is an array of single image object detectors in computer vision literature that we did not train our model on. Due to the time (usually 3-6 days) and GPU resources required to train the model, it is infeasible to train our model using many different types of object detection approaches. Therefore we relied on prior work on similar critical applications to determine what our single image object detector should be. We have cited those prior applications in section 4.2.

It is challenging to explain why a particular deep learning system produces a certain result. A whole academic field has been formed around explainability in deep learning and AI systems. For instance, why the model fails to detect a rip current even though there is a visible rip current to the human observer in the frame?. We attempt to explain the variability in the model's accuracy values by visualizing the input image in the feature space similar to the works of [80] and [81]. As an example, we choose frames 0 and 833 of *rip_21.mp4*. For both frames, there is a visible rip current observable to the human expert. For frame 0, the model detects a rip current (true positive), but for frame 833, it fails to detect a rip current (false negative). By observing the input images in the feature space as shown in Figure 13, we can see that for frame 0, the features resemble a rip current, and for frame 833, the features do not resemble a rip current. We attribute the true positive detection for frame 0 to the strong signal generated by the features. Similarly, we can attribute the false negative of frame 833 to the weak signal generated by its features.

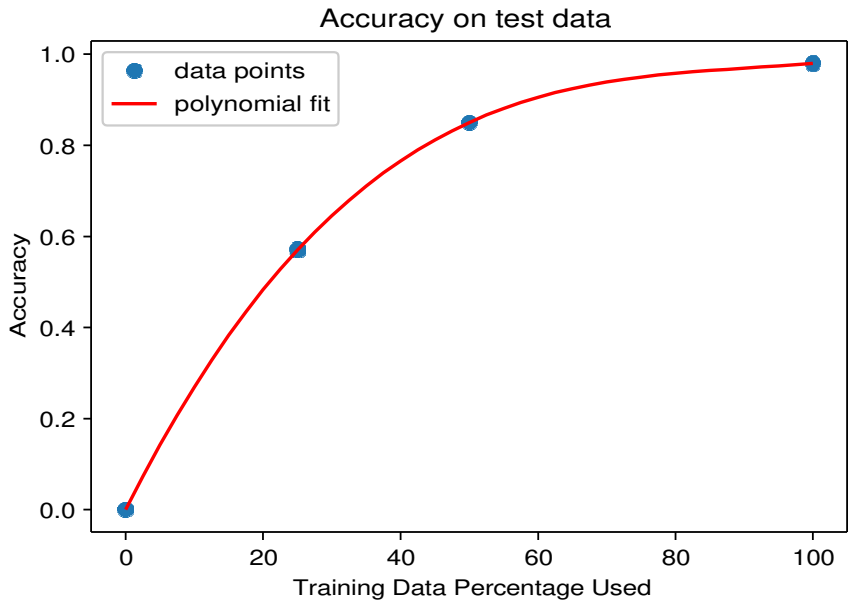

Figure 11: Accuracy for models trained on different percentages of training data.

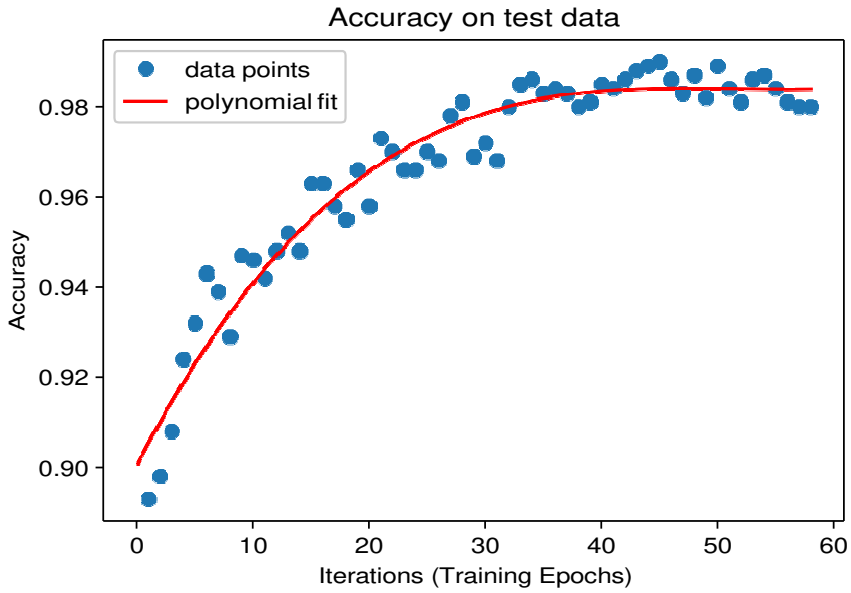

Figure 12: Accuracy for testing data on each training epoch

We also noticed that Maryan et al. [59] did not perform well in non-rip current cases as shown in Maryan[59] column in Table 1. After further study we realized that the test set they used contained almost entirely of images with rip currents. We think that the model proposed by Maryan was not extensively tested on non-rip current images, which accounted for its poor performance for such conditions.

We did not do transfer learning to train our model. Transfer learning is a machine learning technique where the

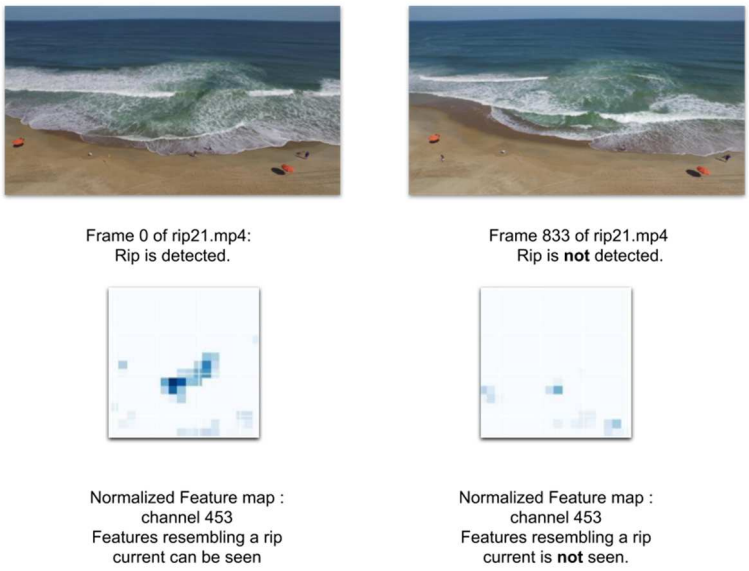

Figure 13: Visualization of frame 0 and frame 833 of *rip*_21.*mp*4 in feature space

parameters and weights learned in one problem are used as the initial values for training another model on a different problem [75]. This strategy can significantly reduce training time and is useful when the two problems belong to similar domains e.g. detecting cars vs detecting tanks. We were not able to exploit this strategy in training our models as we could not find a deep learning model that was trained on a domain somewhat similar to detecting rip currents. Note that the method proposed by Maryan et al. [59] is not a deep learning method and hence not applicable here.

We found it difficult to compare our method with prior work and verify that our model performs well in all conditions previously researched, due to a lack of public data sets on which to verify our results. In order to ensure that future work has a baseline from which to compare, our data sets with thousands of labeled frames are available as part of the supplementary material. Nevertheless our data sets are still limited. The accuracy numbers presented in this paper are correct on this limited data, but almost certainly overstate probable outcomes in real world deployment. We expect that future work will need to collect more examples including less common rip current visual presentation, a greater variety of scales, and a wider array of beach distractors.

Lastly, our work can benefit from a re-examination of the IoU metric employed by our algorithm. Certainly IoU, true positive rate, mean average precision (mAP), and the like are common in computer vision research, but these are usually used in the context of detection based on appearance rather than on behavior. It would be interesting to study how deep learning methods can be trained to recognize rip current behavior using a metric that incorporates a temporal dimension.

7. Conclusion

We present a machine learning approach for identifying rip currents automatically. We use Faster R-CNN and a custom temporal aggregation stage to make detections from still images or videos with higher measured accuracy than both humans and other methods of rip current detection previously reported in the literature. Training data and test data are included in the supplementary materials.

A. Appendix

In order to encourage progress in this domain, both training and test data sets will be made available to the public. Supplementary material including the results to this article can be found online at https://sites.google.com/view/ripcurrentdetection/home

B. Acknowledgements

We thank Ra'Teema Stanley for labelling the data from Miami beach webcams. We also thank Boglárka Ecsedi and Tara Natarajan for collecting some of the training data.

This report was prepared in part as a result of work sponsored by the Southeast Coastal Ocean Observing Regional Association (SECOORA) with NOAA financial assistance award number NA20NOS0120220. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of SECOORA or NOAA.

References

[1] Athira, M., Khan, D.M., 2020. Recent trends on object detection and image classification: A review, in: 2020 International Conference on Computational Performance Evaluation (ComPE), IEEE. pp. 427–435.

[2] Australia, S.L.S., 2019. National coastal safety report. https://issuu.com/surflifesavingaustralia/docs/ncsr2019 .

[3] Barmpoutis, P., Dimitropoulos, K., Kaza, K., Grammalidis, N., 2019. Fire detection from images using faster r-cnn and multidimensional texture analysis, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 8301–8305.

[4] Barron, J.L., Fleet, D.J., Beauchemin, S.S., 1994. Performance of optical flow techniques. International Journal of Computer Vision 12, 43–77. URL: https://doi.org/10.1007/BF01420984, doi:10.1007/BF01420984.

[5] Bowen, A.J., 1969. Rip currents: Theoretical investigations. Journal of Geophysical Research 74, 5467–5478.

[6] Branche, C.M., 2001. Lifeguard effectiveness: A report of the working group. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control URL: https://www.cdc.gov/homeandrecreationalsafety/pubs/lifeguardreport-a.pdf.

[7] Brander, R.W., Drozdzewski, D., Dominey-Howes, D., 2014. "Dye in the Water": A visual approach to communicating the rip current hazard. Science Communication 36, 802–810.

[8] Brannstrom, C., Brown, H., Houser, C., Trimble, S., Lavoie, A., 2015. "You can't see them from sitting here": Evaluating beach user understanding of a rip current warning sign. Applied Geography 56, 61–70. doi:10.1016/j.apgeog.2014.10.011.

[9] Brewster, B.C., Gould, R.E., Brander, R.W., 2019. Estimations of rip current rescues and drowning in the United States. Natural Hazards and Earth System Sciences 19, 389–397.

[10] Brighton, B., Sherker, S., Brander, R., Thompson, M., Bradstreet, A., 2013. Rip current related drowning deaths and rescues in Australia 2004-2011. Natural Hazards and Earth System Sciences 13, 1069–1075.doi:10.5194/nhess-13-1069-2013.

[11] Brocchini, M., Kennedy, A., Soldini, L., Mancinelli, A., 2004. Topographically controlled, breaking-wave-induced macrovortices. Part 1. Widely separated breakwaters. Journal of Fluid Mechanics 507, 289–307. URL: http://www.journals.cambridge.org/abstract_S002211200400878X, doi:10.1017/S002211200400878X.

[12] Buscombe, D., Carini, R.J., 2019. A Data-Driven Approach to Classifying Wave Breaking in Infrared Imagery. Remote Sensing 11, 859. URL: https://www.mdpi.com/2072-4292/11/7/859, doi:10.3390/rs11070859.

[13] Buscombe, D., Carini, R.J., Harrison, S.R., Chickadel, C.C., Warrick, J.A., 2020. Optical wave gauging using deep neural networks. Coastal Engineering 155, 103593. URL: http://www.sciencedirect.com/science/article/pii/S0378383919301243, doi:10.1016/j.coastaleng.2019.103593.

[14] Buscombe, D., Ritchie, A.C., 2018. Landscape Classification with Deep Neural Networks. Geosciences 8, 244. URL: https://www.mdpi.com/2076-3263/8/7/244, doi:10.3390/geosciences8070244.

[15] Castelle, B., Almar, R., Dorel, M., Lefebvre, J.P., Senechal, N., Anthony, E.J., Laibi, R., Chuchla, R., du Penhoat, Y., 2014. Rip currents and circulation on a high-energy low-tide-terraced beach (Grand Popo, Benin, West Africa). Journal of Coastal Research 70, 633 – 638. URL: https://doi.org/10.2112/SI70-107.1, doi:10.2112/SI70-107.1.

[16] Castelle, B., Scott, T., Brander, R., McCarroll, R., 2016. Rip current types, circulation and hazard. Earth-Science Reviews 163, 1–21.

[17] Chen, J.J., Menezes, N.J., Bradley, A.D., North, T., 2011. Opportunities for crowdsourcing research on amazon mechanical turk. Interfaces 5, 1.

[18] Clark, D., Feddersen, F., Guza, R., 2010. Cross-shore surfzone tracer dispersion in an alongshore current. Journal of Geophysical Research (Oceans) 115. doi:10.1029/2009JC005683.

[19] Clark, D.B., Lenain, L., Feddersen, F., Boss, E., Guza, R., 2014. Aerial imaging of fluorescent dye in the near shore. Journal of Atmospheric and Oceanic Technology 31, 1410–1421.

[20] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248–255.

[21] Dérian, P., Almar, R., 2017. Wavelet-based optical flow estimation of instant surface currents from shore-based and UAV videos. IEEE Transactions on Geoscience and Remote Sensing 55, 5790–5797.

[22] Dhillon, A., Verma, G.K., 2020. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence 9, 85–112.

[23] Ducournau, A., Fablet, R., 2016. Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived SST data, in: 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS), pp. 1–6. doi:10.1109/PRRS.2016.7867019. iSSN: null.

[24] Dusek, G., Hernandez, D., Willis, M., Brown, J.A., Long, J.W., Porter, D.E., Vance, T.C., 2019. WebCAT: Piloting the development of a web camera coastal observing network for diverse applications. Frontiers in Marine Science 6, 353.

[25] Dusek, G., Seim, H., 2013. A probabilistic rip current forecast model. Journal of Coastal Research 29, 909–925. URL: <GotoISI>://WOS:000321162100015, doi:10.2112/Jcoastres-D-12-00118.1.

[26] Elgar, S., Raubenheimer, B., Guza, R.T., 2001. Current meter performance in the surf zone. Journal of Atmospheric and Oceanic Technology 18, 1735–1746. URL: <GotoISI>://WOS:000171624800010, doi:10.1175/1520-0426(2001)018<1735:cmpits>2.0.co;2. n/a.

[27] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision 88, 303–338.

[28] da F. Klein, A.H., Santana, G.G., Diehl, F.L., de Menezes, J.T., 2003. Analysis of hazards associated with sea bathing: Results of five years work in oceanic beaches of Santa Catarina state, Southern Brazil. Journal of Coastal Research , 107–116URL: http://www.jstor.org/stable/40928754.

[29] Girshick, R., 2015. Fast R-CNN, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

[30] Gupta, S., 2018. Deep learning performance breakthrough. URL: https://www.ibm.com/blogs/systems/deep-learning-performance-breakthrough/. library Catalog: www.ibm.com.

[31] Haller, M.C., Honegger, D., Catalan, P.A., 2014. Rip current observations via marine radar. Journal of waterway, port, coastal, and ocean engineering 140, 115–124.

[32] Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D., 2018. Advanced deep-learning techniques for salient and category-specific object detection: A survey. IEEE Signal Processing Magazine 35, 84–100.

[33] Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S., 2016. Seq-NMS for video object detection. CoRR abs/1602.08465. URL: http://arxiv.org/abs/1602.08465,arXiv:1602.08465.

[34] Holland, K.T., Holman, R.A., Lippmann, T.C., Stanley, J., Plant, N., 1997. Practical use of video imagery in nearshore oceanographic field studies. IEEE Journal of Oceanic Engineering 22, 81–92.

[35] Holman, R.A., Stanley, J., 2007. The history and technical capabilities of Argus. Coastal Engineering 54, 477–491.

[36] Holman, R.A., Symonds, G., Thornton, E.B., Ranasinghe, R., 2006. Rip spacing and persistence on an embayed beach. Journal of Geophysical Research-Oceans 111. URL: <GotoISI>://WOS:000234999800001, doi:ArtnC0100610.1029/2005jc002965.

[37] Horn, B.K.P., Schunck, B.G., 1981. Determining optical flow. Artificial Intelligence 17, 185–203. URL: http://dx.doi.org/10.1016/0004-3702(81)90024-2, doi:10.1016/0004-3702(81)90024-2.

[38] Houser, C., Trimble, S., Brander, R., Brewster, B.C., Dusek, G., Jones, D., Kuhn, J., 2017. Public perceptions of a rip current hazard education program: Break the Grip of the Rip! Natural Hazards and Earth System Sciences 17, 1003–1024. URL: <GotoISI>://WOS:000404797900001, doi:10.5194/nhess-17-1003-2017.

[39] Hung, J., Carpenter, A., 2017. Applying faster r-cnn for object detection on malaria images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[40] Inch, K., 2014. Surf zone hydrodynamics: Measuring waves and currents. Geomorphological Techniques Chap. 3, Sec 2.3.

[41] Jaiswal, S., Jaiswal, T., 2020. Deep learning approaches for object detection. Artificial Intelligence Evolution , 122–144.

[42] Jiang, G.Q., Xu, J., Wei, J., 2018. A deep learning algorithm of neural network for the parameterization of typhoon-ocean feedback in typhoon forecast models. Geophysical Research Letters 45, 3706–3716. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2018GL077004, doi:10.1002/2018GL077004.

[43] Johnson, D., Pattiaratchi, C., 2004. Transient rip currents and nearshore circulation on a swell-dominated beach. Journal of Geophysical Research: Oceans 109. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003JC001798, doi:10.1029/2003JC001798,arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2003JC001798.

[44] Kamath, C.N., Bukhari, S.S., Dengel, A., 2018. Comparative study between traditional machine learning and deep learning approaches for text classification, in: Proceedings of the ACM Symposium on Document Engineering 2018, pp. 1–11.

[45] Kang, K., Ouyang, W., Li, H., Wang, X., 2016. Object detection from video tubelets with convolutional neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA. pp. 817–825. URL: http://ieeexplore.ieee.org/document/7780464/, doi:10.1109/CVPR.2016.95.

[46] Kennedy, A.B., Brocchini, M., Soldini, L., Gutierrez, E., 2006. Topographically controlled, breaking-wave-induced macrovortices. Part 2. Changing geometries. Journal of Fluid Mechanics 559, 57. URL: http://www.journals.cambridge.org/abstract_S0022112006009979, doi:10.1017/S0022112006009979.

[47] Kim, J.H., Batchuluun, G., Park, K.R., 2018. Pedestrian detection based on faster r-cnn in nighttime by fusing deep convolutional features of successive images. Expert Systems with Applications 114, 15–33.

[48] Kumar, N., Feddersen, F., 2017a. The effect of stokes drift and transient rip currents on the inner shelf. part i: No stratification. Journal of Physical Oceanography 47, 227–241.

[49] Kumar, N., Feddersen, F., 2017b. A new offshore transport mechanism for shoreline-released tracer induced by transient rip currents and stratification. Geophysical Research Letters 44, 2843–2851.

[50] Lima, E., Sun, X., Dong, J., Wang, H., Yang, Y., Liu, L., 2017. Learning and Transferring Convolutional Neural Network Knowledge to Ocean Front Recognition. IEEE Geoscience and Remote Sensing Letters 14, 354–358. URL: http://ieeexplore.ieee.org/document/7829262/, doi:10.1109/LGRS.2016.2643000.

[51] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, Springer. pp. 740–755.

[52] Lippmann, T.C., Holman, R.A., 1989. Quantification of sand bar morphology: A video technique based on wave dissipation. Journal of

Geophysical Research: Oceans 94, 995–1011.
[53] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2018. Deep learning for generic object detection: A survey. arXiv preprint arXiv:1809.02165 .
[54] Liu, Y., Wu, C.H., 2019. Lifeguarding operational camera kiosk system (LOCKS) for flash rip warning: Development and application. Coastal Engineering 152, 103537.
[55] Luo, J.Y., Irisson, J.O., Graham, B., Guigand, C., Sarafraz, A., Mader, C., Cowen, R.K., 2018. Automated plankton image analysis using convolutional neural networks. Limnology and Oceanography: Methods 16, 814–827. URL: https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10285, doi:10.1002/lom3.10285.
[56] Lushine, J.B., 1991. A study of rip current drownings and related weather factors. National Weather Digest , 13–19.
[57] Ma, S., Huang, Y., Che, X., Gu, R., 2020. Faster rcnn-based detection of cervical spinal cord injury and disc degeneration. Journal of Applied Clinical Medical Physics 21, 235–243.
[58] Mahmood, T., Arsalan, M., Owais, M., Lee, M.B., Park, K.R., 2020. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster r-cnn and deep cnns. Journal of Clinical Medicine 9, 749.
[59] Maryan, C., Hoque, M.T., Michael, C., Ioup, E., Abdelguerfi, M., 2019. Machine learning applications in detecting rip channels from images. Applied Soft Computing 78, 84–93.
[60] Meadows, G.A., Grimm, A., Brooks, C.N., Shuchman, R.A., 2015. Remote sensing-based detection and monitoring of dangerous nearshore currents .
[61] Nelko, V., Dalrymple, R., 2011. Rip current prediction in Ocean City, Maryland, in: Rip Currents: Beach Safety, Physical Oceanography and Wave Modeling. CRC Press, pp. 45–58. URL: https://www.academia.edu/1017360/Rip_Current_Prediction_at_Ocean_City_Maryland.
[62] Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on amazon mechanical turk. Judgment and Decision making 5, 411–419.
[63] Papageorgiou, C.P., Oren, M., Poggio, T., 1998. A general framework for object detection, in: Proceedings of the Sixth International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA. pp. 555–562. URL: http://dl.acm.org/citation.cfm?id=938978.939174.
[64] Perrier, G., 2005. Automated rip current detection system. US Patent App. 11/203,771.
[65] Philip, S., Pang, A., 2016. Detecting and Visualizing Rip Current Using Optical Flow, in: Bertini, E., Elmqvist, N., Wischgoll, T. (Eds.), EuroVis 2016 - Short Papers, The Eurographics Association. p. 115. doi:10.2312/eurovisshort.20161155.
[66] Piattella, A., Brocchini, M., Mancinelli, A., 2006. Topographically controlled, breaking-wave-induced macrovortices. Part 3. The mixing features. Journal of Fluid Mechanics 559, 81. URL: http://www.journals.cambridge.org/abstract_S0022112006009918, doi:10.1017/S0022112006009918.
[67] Pitman, S., Gallop, S.L., Haigh, I.D., Mahmoodi, S., Masselink, G., Ranasinghe, R., 2016. Synthetic imagery for the automated detection of rip currents. Journal of coastal research , 912–916.
[68] Postacchini, M., Brocchini, M., Soldini, L., 2014. Vorticity generation due to cross-sea. Journal of Fluid Mechanics 744, 286–309. URL: https://www.cambridge.org/core/product/identifier/S0022112014000445/type/journal_article, doi:10.1017/jfm.2014.44.
[69] Pritchard, D.W., Carpenter, J.H., 1960. Measurements of turbulent diffusion in estuarine and inshore waters. International Association of Scientific Hydrology Bulletin 5, 37–50. URL: https://doi.org/10.1080/02626666009493189, doi:10.1080/02626666009493189, arXiv:https://doi.org/10.1080/02626666009493189.
[70] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, real-time object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 0.
[71] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, pp. 91–99.
[72] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[73] Schmidt, W., Woodward, B., Millikan, K., Guza, R., Raubenheimer, B., Elgar, S., 2003. A GPS-tracked surf zone drifter. Journal of Atmospheric and Oceanic Technology 20, 1069–1075.
[74] Stringari, C.E., Harris, D.L., Power, H.E., 2019. A novel machine learning algorithm for tracking remotely sensed waves in the surf zone. Coastal Engineering 147, 149–158. URL: http://www.sciencedirect.com/science/article/pii/S037838391830228X, doi:10.1016/j.coastaleng.2019.02.002.
[75] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning, in: International conference on artificial neural networks, Springer. pp. 270–279.
[76] Viola, P., Jones, M.J., 2004. Robust real-time face detection. International Journal of Computer Vision 57, 137–154. URL: https://doi.org/10.1023/B:VISI.0000013087.49260.fb, doi:10.1023/B:VISI.0000013087.49260.fb.
[77] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience 2018.
[78] Wu, X., Sahoo, D., Hoi, S.C., 2020. Recent advances in deep learning for object detection. Neurocomputing .
[79] Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y., 2017a. Flow-guided feature aggregation for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 408–417.
[80] Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y., 2017b. Flow-guided feature aggregation for video object detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 408–417.
[81] Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y., 2017c. Deep feature flow for video recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2349–2358.
[82] Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019. Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 .