Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 978451, 11 pages
doi:10.1155/2010/978451

Research Article

An Evaluation of Pixel-Based Methods for the Detection of Floating Objects on the Sea Surface

Alexander Borghgraef,[1] Olivier Barnich,[2] Fabian Lapierre,[1] Marc Van Droogenbroeck,[2] Wilfried Philips,[3] and Marc Acheroy[1]

[1] *Department CISS, Signal and Image Centre, Royal Military Academy, B-1000 Brussels, Belgium*
[2] *INTELSIG Group, Montefiore Institute, University of Liège, B-4000 Liège, Belgium*
[3] *Department Telin, IPI, Ghent University, B-9000 Ghent, Belgium*

Correspondence should be addressed to Alexander Borghgraef, alexander.borghgraef@rma.ac.be

Received 1 July 2009; Revised 16 November 2009; Accepted 12 January 2010

Academic Editor: Frank Ehlers

Ship-based automatic detection of small floating objects on an agitated sea surface remains a hard problem. Our main concern is the detection of floating mines, which proved a real threat to shipping in confined waterways during the first Gulf War, but applications include salvaging, search-and-rescue operation, perimeter, or harbour defense. Detection in infrared (IR) is challenging because a rough sea is seen as a dynamic background of moving objects with size order, shape, and temperature similar to those of the floating mine. In this paper we have applied a selection of background subtraction algorithms to the problem, and we show that the recent algorithms such as ViBe and behaviour subtraction, which take into account spatial and temporal correlations within the dynamic scene, significantly outperform the more conventional parametric techniques, with only little prior assumptions about the physical properties of the scene.

1. Introduction

During the Gulf War of 1990-1991, free floating sea mines proved to be a real threat to shipping in the Persian Gulf. Normally, sea mines are stationary interdiction devices, acting as a deterrent to keep hostile ships from entering strategically important zones. Two types are prevalent: bottom mines, a nonbuoyant type equipped with electromagnetic and acoustic sensors capable of detecting and even identifying passing ships, and the simpler tethered mines, which float just below the sea surface, rely on contact detonators and are anchored to the seabed.

It is the last type, shown in Figure 1, which concerns us. International law dictates that floating mines have to be anchored in place and equipped with a self-disabling mechanism should the anchoring fail. Most nations, even those averse to international law, tend to abide by these regulations, since a floating mine carried by random currents functions as a very inefficient missile and rapidly leaves the conflict area.

However, when the conflict takes place in more confined places with high numbers of ships passing through, the hit probability rises significantly and floating mines can be turned into a poor man's antiship missile. The Persian Gulf is such a region, and a large number of free-floating mines were encountered by Coalition forces.

The effectiveness of this strategy was apparent when US Navy amphibious assault carrier Tripoli was put out of action by a contact mine, after which floating mines were labeled a primary threat to US Navy aircraft carriers. Military operations were considerably slowed down because of it, and risk-averse shipping companies reduced their traffic in the region to a minimum.

Conventional subsurface mines are detected through a variety of sonar techniques, which proved inadequate for spotting floating mines due to the practical matter of downward sensor inclination, but more importantly because of the sea surface presenting a strongly cluttered background in which the small target easily gets lost. The same reasoning applies above water: the target is small and

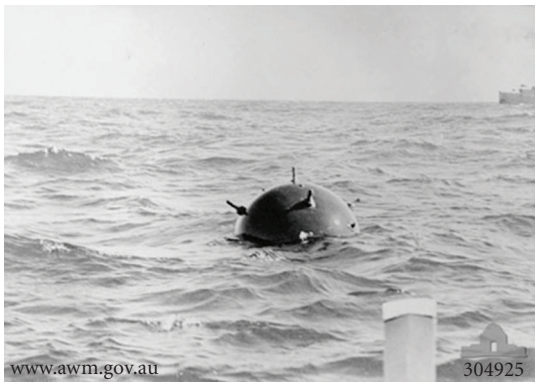


Figure 1: A WWII German floating mine. (Australian War Memorial (AWM) catalog number 304925.)

partially submerged, and the agitated sea surface acts as a dynamic and noisy background hiding the target among the clutter. In practice, the only reliable detection method so far proved to be a human lookout, which significantly complicated operations since every potential contact had to be carefully approached and identified visually. This experience shows that an automated detection system would be beneficial in similar conflicts, and in cleanup operations of the abandoned mines of old conflicts.

Our requirements are for a ship-based collision avoidance system capable of operating day and night and in most weather conditions. The small size of the intended targets and the proven possibility of visual detection by human operator led us to select an infrared video imager as our sensor of choice. Optical image intensifiers were considered but rejected due to their worse performance in bad weather conditions, and their dependence on target colour and thus higher susceptibility to camouflage.

Automated detection of the floating mine is hard, because waves on the sea surface act as a dynamic background which is difficult to distinguish from the foreground mine. Object-based tracking methods such as mean-shift tracking [1] and particle filters [2] can in theory be used to obtain a background motion model, which can then be used to classify object tracks as foreground or background. In practice, they have a hard time dealing with the large quantity and high variability of the dynamic background objects, and with the limited object information present in infrared sequences [3]. Optical flow [4, 5] and block-matching [6] methods have been used with more success. Also, the system needs real-time performance to be useful, which caused us to prefer algorithms operating at the pixel level.

In this paper we show that background subtraction techniques can be used to detect small objects on the sea surface. We take two recent algorithms developed for video surveillance, the ViBe sample-based method [7] and the behaviour subtraction algorithm [8], and apply them to the floating mine problem. These methods differ from classical background subtraction in that they take into account the spatial (ViBe) and temporal correlation (behaviour subtraction) of pixels in the scene when evaluating and updating their background model.

We have adapted the behaviour subtraction training process to take advantage of the expected horizontal invariance of background behaviour, and propose a memory-saving implementation. We evaluate these methods by comparing their performance on a number of video sequences to that of a number of classical background subtraction methods [9–12], including the state of the art in parametric background subtraction [13, 14] and show that both ViBe and behaviour subtraction significantly outperform these methods.

In this paper, we have taken a rather generic approach, involving a minimum of specific prior target information into the model. As a result, targets are simply detected as anomalies with respect to the background model, which allows for many possible applications outside the domain of mine warfare. Classification of small-sized anomalies in the dynamic background of the sea surface has many civilian applications as well. Debris, containers, swimmers, small boats, and even the snorkels of semi-submersible smuggling vessels and subsurface sandbanks all appear as anomalies in the background behaviour, leading to a wide range of applications such as salvaging, search-and-rescue operation, lifeguard assistance, and coast-guard operations.

In Section 2, we describe the physical characteristics of a typical scene containing a target on the sea surface, as observed by a ship-mounted thermal camera. In Section 3, we provide an overview of three classes of background subtraction techniques, with a more detailed description of the (Extended Gaussian Mixture Model) EGMM [13, 14] and ViBe methods. Section 4 describes a more complex technique called behaviour subtraction, which includes temporal information into the background model. Finally in Section 5, we apply five different algorithms to a set of IR video sequences of targets floating on the sea surface, and evaluate their performance in detecting the target.

2. Physics and Geometry of the Scene

2.1. The Scene. The requirement of a ship-mounted sensor places the sensor around 10 m above the sea surface, which given a minimal detection distance of 500 m leading to very sharp observation angles of 1° or less. This number lies significantly below the Brewster angle, precluding the use of polarization filters as a means to improve detection probability.

The objects we wish to consider vary in size and shape. Mines are typically semispherical objects of between 1 and 3 m diameter or approximately 2×1 m cylinders. Standard sized shipping containers are $6 \times 2.5 \times 2.5$ m beams, putting an upper bound on the size range we wish to detect. Assuming that half of the object's volume is submerged, we obtain an observable solid angle of 3.1–30 μsr at range 500 m, going to 0.8–7.5 μsr at 1000 m distance.

Sea surface behaviour consists of a superposition of random sinusoidal surface waves and is mainly driven by the wind velocity. The statistical characterization of these waves was first attempted in the 1950s [15], culminating in

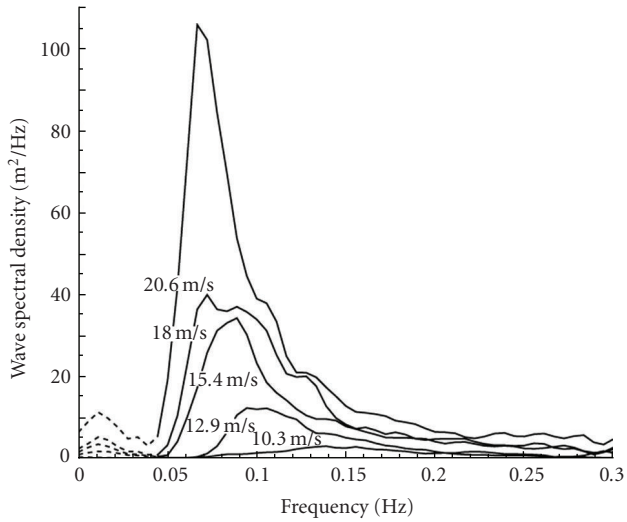


Figure 2: The Pierson-Moskowitz spectral distribution of a fully developed sea at different wind speeds [16, 18].

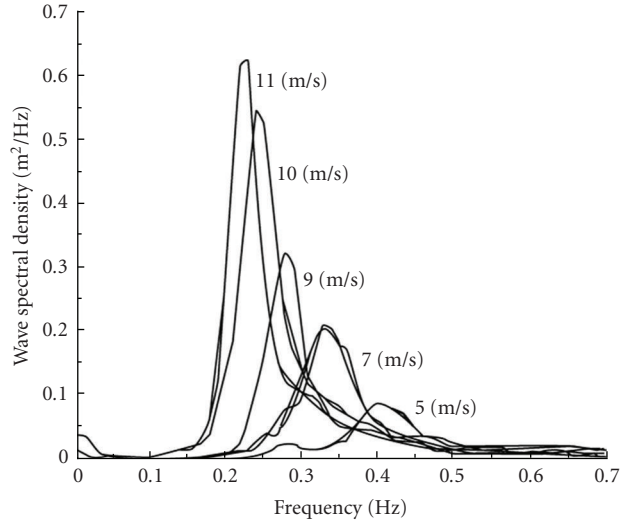


Figure 3: The JONSWAP spectral distribution of a fully developed sea at different wind speeds [17, 18].

a number of empirical models for the spectral distribution of a fully developed sea (meaning in a state of equilibrium with the driving wind), including the model by Pierson and Moskowitz [16], and the later (Joint North Sea Wave Observation Project) JONSWAP model derived from extensive buoy measurement data [17]. These models provide a distribution of ocean wave frequencies with the wind speed as its sole parameter, as shown in Figures 2 and 3 for different wind speeds.

While translating these models directly into a prior model for scene behaviour would be impractical, they do provide us with the valuable insight that the waves in the scene form a coherent behavioural system, which can be estimated or learned, and used to classify scene objects as belonging to the sea surface background, or as outlier objects belonging to the foreground. Also, the spatial invariance

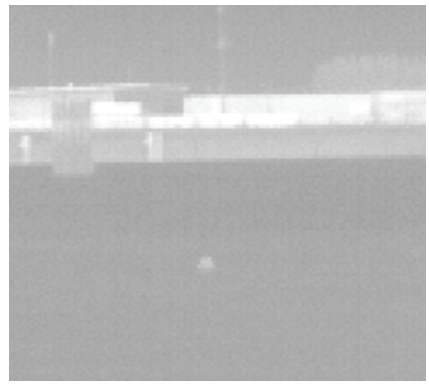


Figure 4: Floating target on a perfectly flat water surface.

of the models points towards the usefulness of pixel-based (as opposed to object-based) sampling of the dynamic background and can be used to derive the invariance of behaviour along the horizontal axis, which we use later on in our implementation of the behaviour subtraction algorithm.

2.2. Contrast. Contrast in the video sequences taken by our MWIR (3–5 μm) and LWIR (8–12 μm) cameras amounts to a difference of temperatures between sea surface and target pixels. We assume the worst case scenario that the target has reached thermal equilibrium with the sea water in which it has been submerged for some time. This is a realistic proposition for most floating mines, with the exception of mines still attached to their mooring (brought above surface by a low tide) or those stabilized by ballast, and this during the daytime. In these cases, the target floats same side up, allowing that side to be heated by solar illumination, which results in a temperature gradient with a maximum higher than the surrounding water. Clearly, nonmine targets such as swimmers, small boats, and semi-submersible snorkels are not at thermal equilibrium, simplifying detection.

Paradoxically, a mine at sea temperature is actually quite easy to detect against the background of a flat sea surface. Because our sensor is so close to the surface, the direction of observation and the surface normal are close to perpendicular, leading to a low observable emissivity, so the blackbody radiation emitted by the sea surface will not feature in the signal received at the sensor. Low emissivity implies high reflectance, so what will be received at the sensor is the reflection of the sky's thermal emissions near the horizon, which are typically at far lower temperatures than those of the surface.

This makes floating object detection on a flat sea a nearly trivial case, as seen in Figure 4, and actually shows an advantage of using a ship-based sensor over an airborne one. The problem arises when a seastate of 2 or higher is attained. In this case, the waves do allow for observation angles close to 45°, allowing for seawater thermal emissions to become observable. This fills the image with a large amount of surfaces of the same intensity and similar scale as the observed target, causing great ambiguity with regard to the distinction between foreground and background.

2.3. Temporal Characteristics. The similarity in intensity-based contrast between the target and the dynamic background shows a necessity to look at other determinant characteristics. As noted in the introduction, a human observer is the current state of the art in the detection of small floating objects. An observer can easily spot the vertical bobbing motion of a floating target and distinguish it from the waves' propagating motion. A more rigorous look into the physical situation shows that while waves are basically vertical oscillations of the water surface, the visible part of the wave is its front, which moves in a linear motion over the sea surface (though foam and the breaking of waves at high sea states can cause local distortions of this linearity).

The floating target on the other hand will show a different movement pattern. For one, it will follow the movement of the sea surface and will therefore exhibit a vertical oscillation with the frequency of the passing waves. Also, currents can cause it to move in a different direction from the waves, and when propulsed by the waves, its drag in the water will have it move at a slower speed than the passing waves.

3. Background Subtraction

3.1. Introduction. Pixel-based background subtraction techniques are among the most important and widely used tools in video analysis [9–11]. It involves a class of change detection algorithms in which a per-pixel statistical model is estimated for the background. This background model is then used to classify the incoming video stream's pixels as foreground or background. Only the foreground pixels are retained, thus providing a change image which can be used for various kinds of object detection, classification, and tracking purposes.

The naive example of background subtraction involves thresholding the difference image $D_t(\vec{x}) = |I_t(\vec{x}) - B(\vec{x})|$ between the current frame $I_t(\vec{x})$ and a static background image $B(\vec{x})$, and this for all pixel positions $\vec{x}$. From here on, we will refrain from explicit mentioning of pixel positions and assume that all operations are performed on each pixel. Applying a threshold to the difference image D_t way we obtain a binary mask M_t:

$$M_t \longleftarrow D_t > \Theta. \tag{1}$$

This approach immediately reveals the key difficulties of background subtraction. Choosing the background image, preferably in an automatic manner is the first problem. More importantly, it is clear that the background is not static except in the trivial cases (e.g., in the blue screen techniques used for movie special effects). Changing illuminations, parked cars leaving their spot (causing holes to appear in the static background), waving tree leaves, and moving shadows are typical examples found in video surveillance.

An extensive list of problems can be found in [19], but we will restrict ourselves to finding solutions for those relevant to our mine detection applications. These are

(i) slow illumination changes: less important due to the choice for IR, though temperatures will change during the day,

(ii) camera movement: partially compensated through stabilized gyroscopic mounts,

(iii) high-frequency background movement: caused by waves, referred to as the dynamic background in this paper.

Realistic background subtraction algorithms will use statistical pixel properties as selection criteria, obtained during an initial training phase and continuously updated throughout their operation (also known as background maintenance). We group these algorithms into three classes: basic methods, parametric methods, and sample-based methods.

3.2. Basic Methods. Basic methods use simple statistical measures of the video data to describe the background. These methods allow for fast calculations and are easy to understand. Typical pixel characteristics are the mean or median of the pixel's recent history. This requires that a history of n frames should be kept in memory, which can lead to high memory requirements.

A commonly used solution for reducing memory usage is provided by calculating a running mean [10], using a forgetting factor α. This factor determines how long the influence of old data remains present in the model and has a typical value of around 0.95. In this method, at time t, the background model for the next frame B_{t+1} is calculated from the current background model and the current frame. By this way, no history needs to be kept in memory

$$B_{t+1} \longleftarrow \alpha B_t + (1-\alpha) I_t. \tag{2}$$

In all these methods, foreground/background classification is done by thresholding $|I_t - B_t|$. This illustrates one of the problems with these methods: there is no automatic way provided for determining the threshold. Other weaknesses are the lack of correlation between neighbouring pixels, and the inclusion of foreground pixels in the background model. A solution for the last problem can be obtained by masking foreground pixels from the running mean update. Still, the many weaknesses of the basic methods tend to relegate them to the role of preprocessing steps. We will use the running mean later on when implementing the behaviour subtraction algorithm.

3.3. Parametric Methods. A different approach assumes a known type of probability density function for the background model and estimates the distribution parameters from the available historical data. A simple example is the running Gaussian mean as used in the Pfinder algorithm [20]. Here, a single Gaussian distribution with parameters μ and σ is fitted to a training set, giving an initial background model. This model is then updated via a running mean and running variance method:

$$\begin{aligned} \mu_{t+1} &\longleftarrow \alpha\mu_t + (1-\alpha) I_t, \\ \sigma^2_{t+1} &\longleftarrow \alpha\sigma^2_t + (1-\alpha)(I_t - \mu_t)^2. \end{aligned} \tag{3}$$

This method is fast and requires little memory. Also, it allows for a justifiable threshold based upon the distributions standard deviation. However, it is limited to unimodal backgrounds.

Backgrounds are rarely describable by a unimodal model. The sea surface in infrared, for example, has at least two modes: the reflected sky radiance and the blackbody emission by wavefronts. Most practical parametric models will hence use a Gaussian mixture model to describe the background.

We consider the (Extended Gaussian Mixture Model) EGMM [13, 14] to represent the state of the art in parametric background subtraction techniques. This is an adaptation of the adaptive GMM algorithm first described by Stauffer et al. [21]. A GMM describes both background and foreground as a mixture of M separate Gaussians with mean and standard deviation μ_i and σ_i, and mixing weight ω_i. This leads to the following probability distribution for pixel intensity I_t:

$$p(I_t \mid \mathrm{BG} \cup \mathrm{FG}) = \sum_{i=1}^{M} \omega_i \mathcal{N}(I_t; \mu_i, \sigma_i). \tag{4}$$

Parameters are estimated from image history over a time window of T frames, or by approximation through an update process similar to the running Gaussian described in (3). Here, incoming data is weighed by a learning factor $\lambda \approx 1/T$, and assigned to the nearby mode with the largest mixing weight ω_i by a binary ownership label o_i which equals 1 for pixels belonging to the ith nearby mode, and 0 otherwise. This leads to the update equations

$$\omega_i \longleftarrow \omega_i + \lambda(o_i - \omega_i), \tag{5}$$

$$\mu_i \longleftarrow \mu_i + o_i \frac{\lambda}{\omega_i}(I_t - \mu_i), \tag{6}$$

$$\sigma_i^2 \longleftarrow \sigma_i^2 + o_i \frac{\lambda}{\omega_i}\left(\left(I_t - \mu_i\right)^2 - \sigma_i^2\right). \tag{7}$$

Modes are considered nearby if their Mahalanobis distance from the pixel intensity falls within three standard deviations. If no modes are considered nearby, a new Gaussian component with weight λ is added to the distribution centered on the new data point and with a large σ_0. If a set maximum number of components is exceeded, the lowest weighted ones will be removed from the distribution.

In this process a foreign object appearing in the scene will add a new low-weight Gaussian initialized with a high variance to the mixture. This leads to the conclusion that background-foreground segmentation can be achieved by selecting the mixture of the N components of the highest weight to variance ratios as background model, and the remainder as foreground.

Because of this acceptance criterion, the GMM contains the intrinsic assumption that background pixels are of low variance. This will prove to be a weakness when the GMM is applied on IR video footage of a rough sea surface, a context in which the background variance is actually quite high.

The EGMM by Zivkovic et al. improves on this by using normalized mixture weights ω_i to define an underlying multinomial distribution describing the probability that a sample pixel belongs to the ith component of the GMM. This way ω_i is determined by the ratio of the number of samples assigned to component i over T frames:

$$\omega_i = \frac{n_i}{T} = \frac{1}{T}\sum_{j=1}^{T} o_i^{(j)}. \tag{8}$$

This estimate is further improved upon by introducing prior knowledge in the form of the conjugate prior of the multinomial distribution. This is expressed in the update (5) as a negative weight $-c$ imposing a minimal amount of evidence required from the data before a component can be allowed to exist. A new weight update is defined

$$\omega_i \longleftarrow \omega_i + \lambda(o_i - \omega_i - c_T) \tag{9}$$

with $c_T = c/T \approx \lambda c$. After each iteration, the weights must be normalized to ensure a proper probability distribution.

This method can adapt to changing circumstances and a wide range of anomalous foreground objects being introduced in the scene, while still retaining control over the number of Gaussians in the mixtures. A weakness is the arbitrary parameter N: the number of modes included in the background model.

3.4. Sample-Based Methods. Sample based methods take the Monte Carlo approach of using the observed samples directly as an approximation of their generating distribution, instead of fitting them to a parametric distribution. This leads to a greater noise resilience, an ability to model non-Gaussian distributions, and a rapid reaction capability to high-frequency events in the background, at the price of having to keep a history of samples in memory.

Calculating the probability from a set of samples $\mathcal{X} = \{I^{(1)}, I^{(2)}, \ldots, I^{(N)}\}$ is typically done by using a kernel density estimation (KDE) procedure [22], also known as a Parzen window method [23]. We assume that the set $\mathcal{X}$ has been initialized by a training set of background data and will afterwards be updated with samples restricted to those classified as background. This allows us to calculate the background probability density function by evaluating the weighted distance between the pixel value I_t and all samples $I^{(j)}$ from $\mathcal{X}$ inside a volume of size h^d around it (with d being the dimensionality of the pixel value, in our case of monochrome infrared images, $d = 1$):

$$p(I_t \mid \mathrm{BG}) = \frac{1}{Nh^d}\sum_{i=1}^{N} K\left(\frac{\left\|I_t - I^{(j)}\right\|}{h}\right). \tag{10}$$

The kernel function K we use is a uniform step function, effectively counting k, the number of samples $I^{(j)}$ inside the volume

$$p(I_t \mid \mathrm{BG}) = \frac{k}{Nh^d}. \tag{11}$$

This model allows for two easily implemented approaches toward classifying an incoming pixel value as foreground or background. One approach uses a fixed

window size h and accepts the pixel as background if $k > k_{\text{thr}}$. Another approach is the balloon estimator, in which the window size h is increased until a minimum number of samples from $\mathcal{X}$ are covered. In this case the acceptance criterion depends on the window volume: $h^d < V_{\text{thr}}$.

Updating a sample-based model is simple: a pixel classified as background is added to the sample set $\mathcal{X}$, and the oldest samples in the set are removed. Still, some methods split their sample set into two subsets: one for describing long-term background phenomena, and one for dealing with short-term, high frequency change in the background.

The ViBe [7] method we used, implements a fairly simple kernel density estimator but uses an innovative approach to the model update by including spatial information. Classification of the incoming pixel is done by counting the samples in a fixed-size window around it, with the threshold being 2 samples out of a set of 20. The sample set is not divided in a long- and short-term memory, but the bias of single sample set methods towards recent data is alleviated by a random replacement update scheme. To ensure spatial correlation, when data is inserted into a pixel model, it is also added to a randomly selected neighbouring pixel model. This allows for the spatial correlation between pixels to be taken into account by the model, greatly improving the method's ability to deal with noise and small camera movements. One should note that the ViBe algorithm does not need to be trained as it can be instantaneously initialized using a single image.

4. Movement-Based Methods

4.1. Introduction. The background subtraction methods described above are capable of extracting anomalous objects, even from a moving background. However, they still have trouble with the worst case mine detection scenario we assumed, where a high sea state causes the scene to be filled with moving objects of similar scale, shape, and intensity as the foreground target.

Inherently, all anomaly detection methods make trade-offs regarding computational cost and memory usage, trying to maintain an up-to-date background model with a description length as small as possible. Algorithms such as EGMM discard both temporal and spatial information, retaining only image intensity statistics; whereas the sample-based ViBe method introduces some spatial information.

Given the tendency of floating mines to converge to thermal equilibrium with the surrounding sea water, even spatially correlated intensity statistics will not suffice. We stated that the object's motion provides a better detection characteristic, and in the previous research we showed that the short-term motion information provided by an optical flow algorithm was capable of detecting floating objects [4, 5]. Still, this method, which includes both spatial and temporal information, proved to be slow and restricted to the extrema of the objects motion. Therefore we intend to look at longer-term motion information for more reliable detection.

When looking at long-term motion information, keeping track of spatial information becomes difficult, since the background waves will traverse across the entire image. This would require an object-based approach, in which waves are identified as such and tracked throughout the video sequence. In a confused sea state, it is hard to figure out which wave is which, and intersecting wave trains can seem to break up or merge, which seriously complicates tracking them. In this paper, we restrict ourselves to a low-level, pixel-based approach, meaning that spatial information will be lost, and that detection will have to be based upon temporal characteristics.

4.2. Behaviour Subtraction. An interesting approach can be found in the behaviour subtraction method described by Jodoin et al. [8]. This generic technique was developed for the detection of abnormal behaviour in surveillance footage and is not tied to the physical properties of the sea surface scene. Interesting is also that Jodoin et al. demonstrate their algorithm on video sequences of water surfaces, in which the regular surface waves are included in the background; whereas extraordinary events such as ripples from a stone thrown in the water, or a passing boat, are detected as abnormal events.

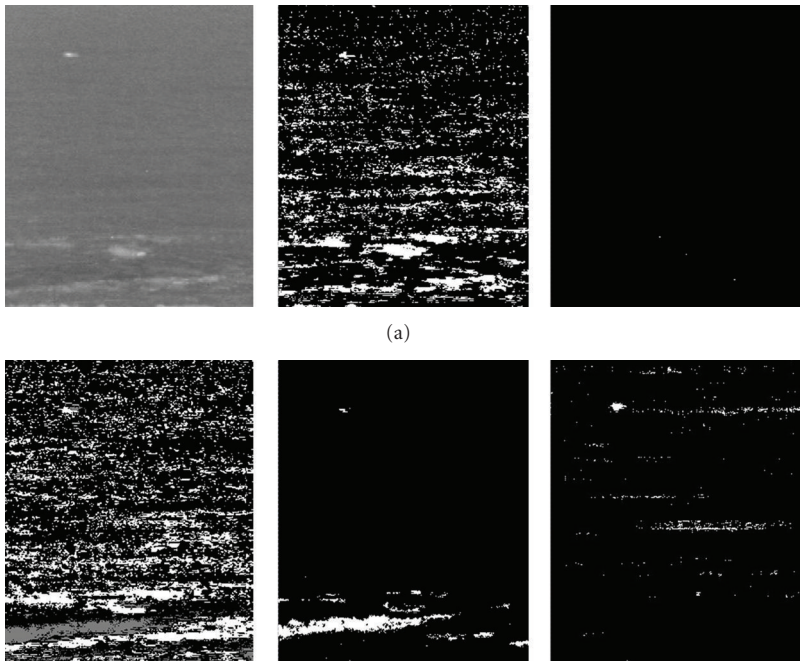
The behaviour subtraction method functions by monitoring the level of scene activity in a sliding time window of length W, and by comparing this observed activity to a background behaviour model derived from a training sequence $\hat{I}$. We explained before that this sort of approach has considerable memory and computational requirements, making real-time implementation difficult. Jodoin et al. solve this problem by significantly reducing the information processed on two levels.

For one, frames are compared to a conventionally obtained background model b (which we calculated using a running mean algorithm). This preprocessing step transforms an integer frame I_t to an image consisting of binary activity labels, thus significantly reducing the required storage space

$$L_t = |I_t - b| > \theta. \tag{12}$$

The second reduction occurs when the label sequence L_t is evaluated in time window of length W. Jodoin et al. reduce the dimensionality of this 3D space-time volume by defining "behaviour descriptors," compressing the dynamics of a pixel into a single value. One behaviour descriptor proposed is the maximum activity descriptor. In this case, an excessively high level of activity is considered to be a sign of abnormal behaviour. Therefore, a background behaviour image B is defined as the maximum level of activity encountered in any of the windows throughout the training sequence $\hat{I}$ of length N:

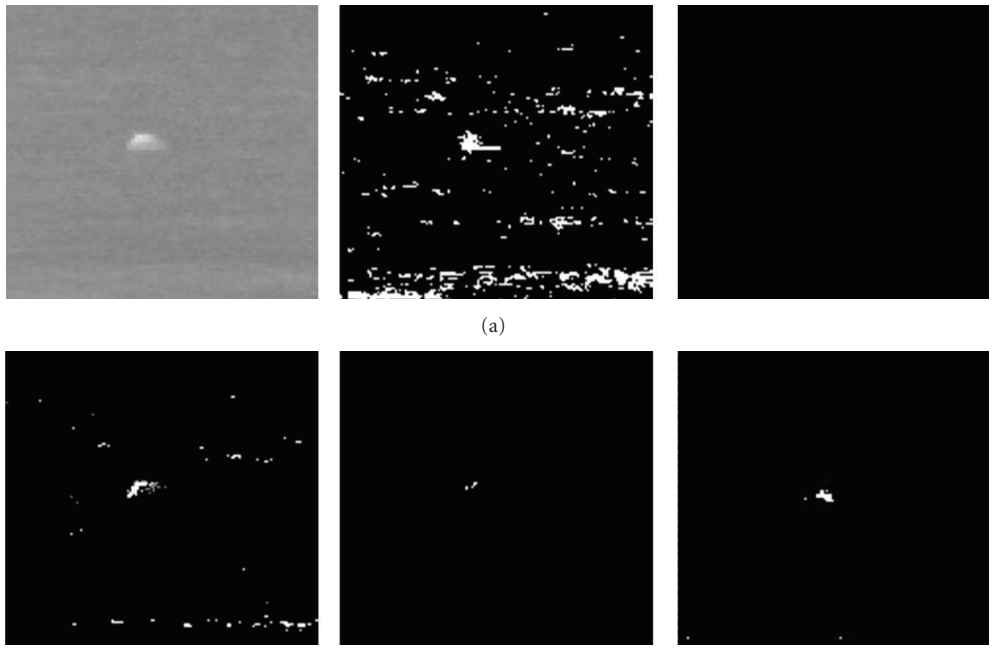
$$B = \max_k \left[\sum_{\tau=k-W+1}^{k} \hat{L}_\tau \right], \quad W \le k \le N. \tag{13}$$



(a)

(b)

Figure 5: MW01 sequence in the 3–5 μm band. From left to right, (a): original, mean, gmm. (b): EGMM, ViBe, behaviour subtraction.



(a)

(b)

Figure 6: MW03 sequence in the 3–5 μm band. From left to right, (a): original, mean, gmm. (b): EGMM, ViBe, behaviour subtraction.

With this background image obtained, the incoming video stream can be evaluated through comparing B to the observed behaviour image v_t at time t:

$$v_t = \sum_{\tau=t-W+1}^{t} L_\tau. \tag{14}$$

This comparison can subsequently be used to detect abnormal behaviour, for example, by defining a "negatives-to-zero" distance function

$$D_t = v_t - \lfloor B \rfloor_0, \tag{15}$$

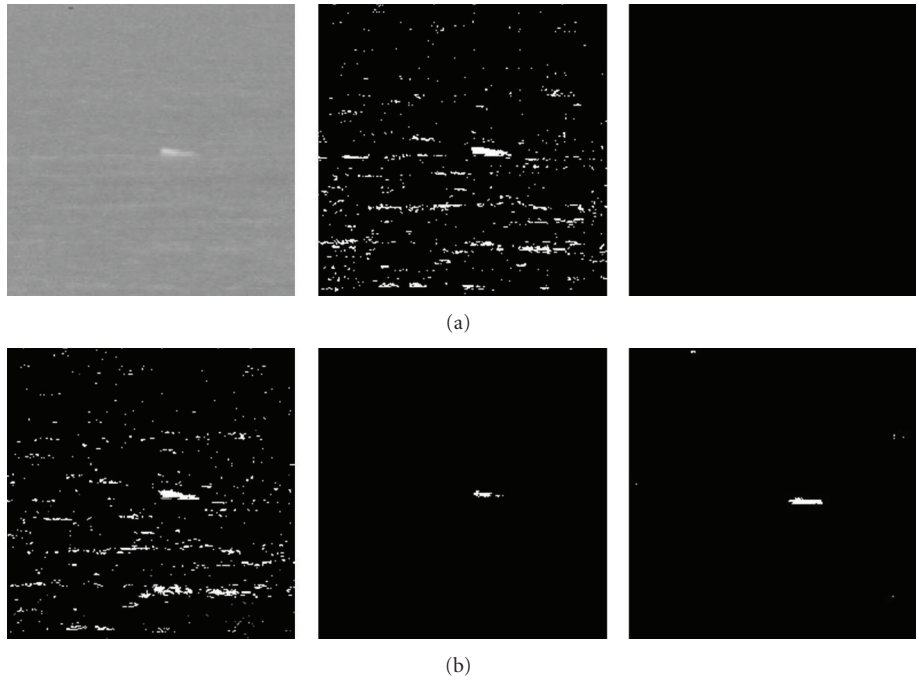


Figure 7: MW05 sequence in the 3–5 μm band. From left to right, (a): original, mean, gmm. (b): EGMM, ViBe, behaviour subtraction.

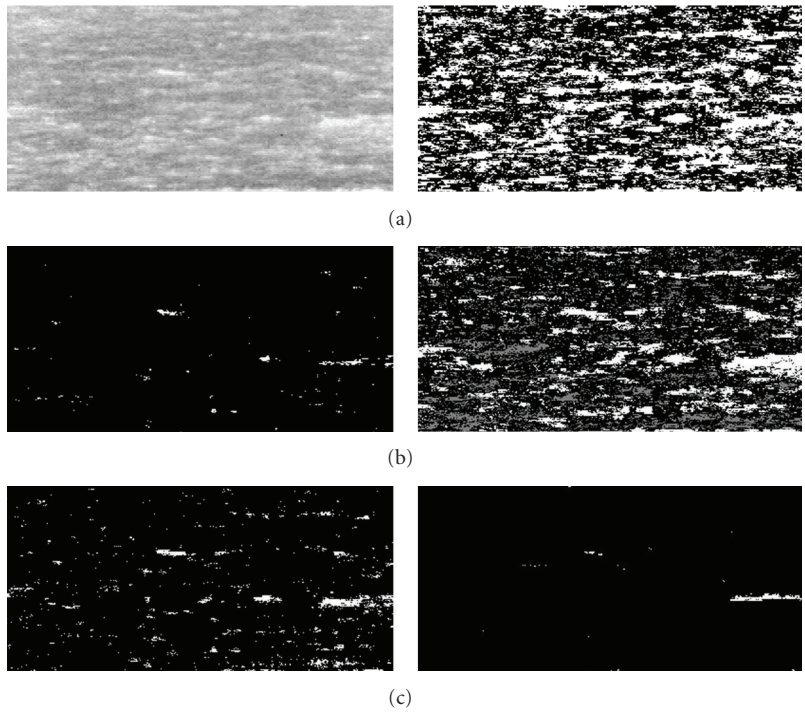


Figure 8: MW08 sequence in the 3–5 μm band. From left to right, (a): original, mean.(b): gmm, EGMM, (c): ViBe, behaviour subtraction.

where $\lfloor a \rfloor_0 = 0$ if $a < 0$. Similarly to (1), a binary mask M_t can be obtained by thresholding this distance function, effectively subtracting the background behaviour and detecting foreground objects.

Realistic assessment of background behaviour requires a fairly large time window W to be chosen. In [8] the authors use $W = 100$ for the maximum activity descriptor. The activity images' binary nature significantly mitigates this large storage requirement, but a conventional high-level implementation still suffers from the number of required memory access operation. A solution can be found in the following writing of (14):

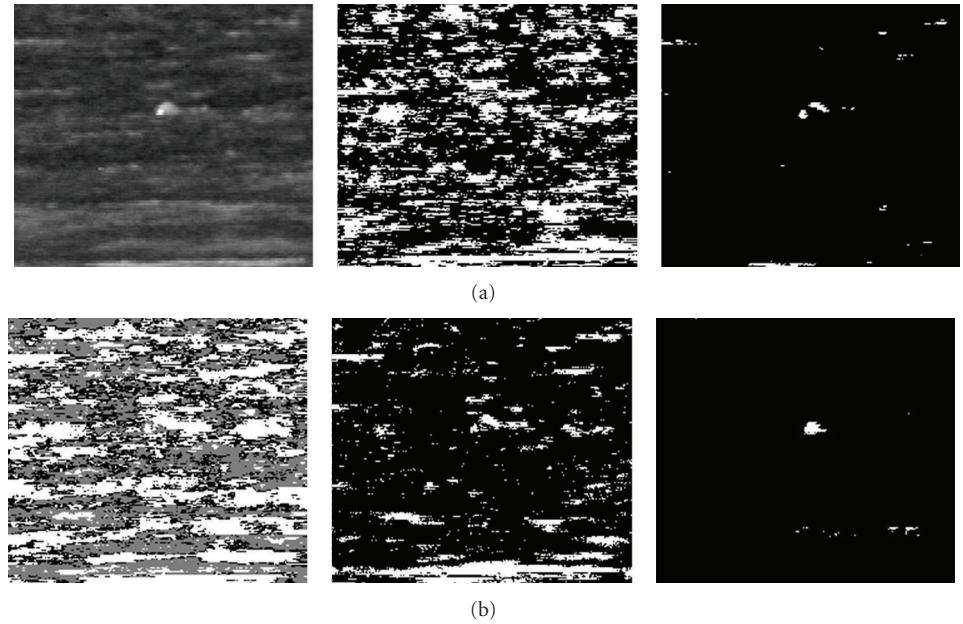


FIGURE 9: MW33 sequence in the 3–5 μm band. From left to right, (a): original, mean, gmm. (b): EGMM, ViBe, behaviour subtraction.

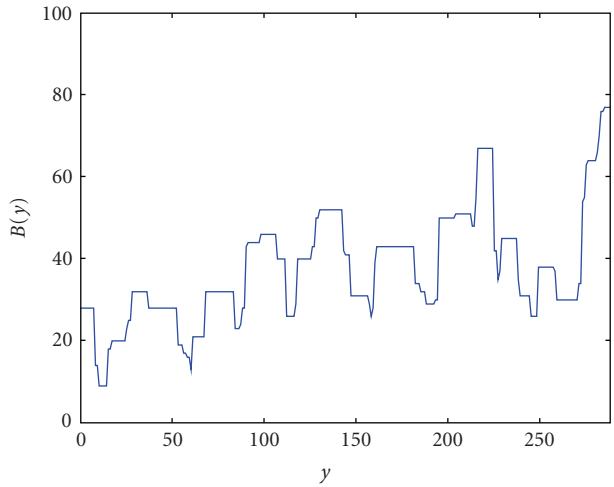


FIGURE 10: Maximum activity descriptor B for MW01 trained on a vertical line (parallel to the y-axis) in the sequence, showing clearly the gaps in the model due to short training sequence. This causes the horizontal streaks of false positive detections by the behaviour subtraction algorithm in Figure 5.

$$v_t = v_{t-1} - L_{t-W} + L_t. \tag{16}$$

Here we see that only 3 images have to be kept in memory at any iteration: the first one of the previous time window L_{t-W}, the previous activity descriptor v_{t-1}, and the current activity image L_t. This keeps memory requirements low for large W and reduces the amount of redundant calculation during summation.

5. Experiments

We used six infrared video sequences of floating test targets taken in various environmental conditions to evaluate the algorithms described above. These sequences are between 300 and 2000 frames long, have been taken in coastal waters using stationary thermal cameras, and have been annotated with groundtruth data for the purpose of detection performance evaluation.

On these measurements we tested a number of pixelbased background subtraction methods described above. We used a running mean, a simple GMM, the extended EGMM algorithm, the ViBe sample-based method, and an adaptation of the behaviour subtraction method. We obtained precision and recall values for the last three algorithms, as seen in Table (1). Precision represents the probability that a detected pixel belongs to a target;whereas recall represents the probability that a target pixel will be detected by the algorithm. This can be expressed in terms of true and false positives (tp, fp) and of false negatives (fn) by the following formula:

$$\begin{aligned} \text{Precision} &= \frac{\text{tp}}{\text{tp} + \text{fp}}, \\ \text{Recall} &= \frac{\text{tp}}{\text{tp} + \text{fn}}. \end{aligned} \tag{17}$$

For the behaviour subtraction implementation, we use the output of the running mean filter to determine the initial background. We also used the prior assumption of spatial invariance in sea surface statistics, which translates into a horizontal invariance of the background behaviour for the scene. This assumption holds in the absence of land, ships, or direct solar glint in the frame. This allowed us to train the behaviour model on a single vertical line of the image, instead of on the entire image. Furthermore we used the

maximum activity descriptor, on a time window of 100 frames and a $\theta = 10$.

When looking at the qualitative results in Figures 5–9, we see ViBe and behaviour subtraction clearly outperforming the parametric models. The classic GMM has few false positives but fails to detect the target in all but one case (video sequence MW33, Figure 9) in which the target intensity is clearly much higher than its surroundings. EGMM, which assigns multiple foreground classes, has the inverse problem: it gets drowned in false positives, only performing reasonably well in one sequence (MW03, Figure 6). We also see this in the quantitative analysis: while EGMM has relatively high recall values, indicating successful detection of the targets, its low precision shows that extracting these targets from the cloud of false positives will be difficult.

It is peculiar that the recall values in Table 1 are all fairly low, which would indicate a low probability of target detection, yet when we look at the images, the targets are clearly visible in the detection results. The reason for this is that our validation is pixel-based, and none of the algorithms registers the full target, but only parts of it, so a number of pixels in the groundtruth data remain undetected. This problem can be avoided by applying post-processing methods, and by, for example, defining detection by coinciding bounding boxes. We chose not to do this, as it would lead us away from the pixel-based paradigm in which the evaluated algorithms were defined.

The sample-based ViBe algorithm performs better overall than the parametric models, though it still has its limitations. It performs exceptionally well in the sequences where there is a significant thermal contrast between target and background (MW01, MW03, and MW05: Figures 5–7), resulting in the heighest precision of all three algorithms, and succeeds in providing a learned model capable of dealing with a very dynamic background. However, when the sea state worsens, and the thermal distinction between the target and the breaking waves diminishes, ViBe's performance weakens (Figures 8 and 9). This was to be expected: ViBe only models the background intensity, which means that it does not possess the temporal information required to distinguish between background and foreground.

Behaviour subtraction on the other hand renders the best performance of all three for high sea state, achieving both high precision and recall in these sequences. It manages to classify all but a few exceptional waves as background, and it identifies the target's behaviour as anomalous in all examples, though only in a part of the target's composing pixels. However, in Figure 5, we show a peculiar failure mode of the algorithm (which performs better later on in the same sequence). We see several series of false positives stringed along horizontal lines. Upon closer inspection, this was caused by the gaps in the behaviour model B which was trained along a vertical line in the image sequence. This lack of smoothness as seen in Figure 10 allows for horizontal zones where the behaviour threshold is lower than in the surroundings, allowing for the strings of false positives to occur. While behavioural variation along the vertical axis is expected due to the difference in range, there is no physical reason for these gaps to occur. This leads us to conclude that smoothness should be imposed on the model, either through longer training times, or by application of a smoothening function to the model.

Table 1: Sequence average precision-recall pairs for behaviour subtraction, ViBe, and EGMM.

	Behaviour	ViBe	EGMM
MW01	.29;.40	.30;.42	.27;.66
MW03	.36;.20	.90;.20	.34;.29
MW05	.25;.62	.86;.49	.09;.29
MW08	.77;.79	.69;.75	.69;.57
MW33	.46;.71	.30;.49	.29;.67

Finally, we would like to indicate a point where all algorithms failed. Sequence MW01 contains two targets, one spherical target in the upper left corner, and a cylindrical one lying in the surf mid-front. None of the algorithms, including behaviour subtraction, succeed in detecting this target, which is the reason for the low precision values measured.

6. Conclusions

In this paper we described the problems involved in the automatic detection of small floating targets on the sea surface, and this in the context of the detection of drifting mines. We described the state of the art in the domain of pixel-based background subtraction algorithms, including the innovative ViBe method, and compared this with the more complex behaviour subtraction method which includes temporal information to address the specific problems of the dynamic background provided by the sea surface.

The sample-based ViBe algorithm performs significantly better than the parametric methods we applied to our test sequences, which is particularly interesting given its complete lack of prior assumptions regarding the scene. In rough seas, the method performs less ideally due to the great similarity between waves and target. While the algorithm could be adapted to incorporate these waves into the background model, this would not allow it to detect targets at thermal equilibrium with the ocean. On the other hand, the algorithm would be exceptionally fit to the detection of targets radiating at a different temperature than the surface, such as swimmers, shipwreckees, small boats, or a semisubmersible's exhaust.

The behaviour subtraction method outperforms all of the background subtraction algorithms presented here, because of its inclusion of temporal information in the model. It deals better with heavy seas than the others and has no trouble detecting targets at equilibrium with their surroundings. This comes at the cost of a heavier computational load, and the need for a long training sequence, which needs to be updated over time. The method shares with ViBe an interesting lack of prior assumptions, allowing it to be applied to a wide range of applications.

We have shown that pixel-based masking techniques such as background and behaviour subtraction can be used to detect anomalous objects in the very dynamic environment provided by an agitated sea surface. While

not suited as stand-alone detection algorithms in all cases, these methods can provide regions of interest with sufficiently high confidence to allow higher-level, object-based classification methods to use them as prior input. In future research we intend to evaluate the alternative "average activity descriptor" described in [8], to obtain ground-truth data for a quantitative validation of the algorithms' detection performance. We also need to improve upon the training procedure for the behaviour subtraction method, guaranteeing a smooth model and allowing for continuous update. Behaviour subtraction also does not use spatial correlation the way ViBe does, which indicates an interesting approach to improving the method. Lastly, since theoretical models of the sea surface describe a spectral distribution, we would like to look into activity descriptors which makes explicit use of frequency information.

Acknowledgments

This research resides within the framework of the MRN06 project sponsored by the Belgian Ministry of Defence and was conducted at the Royal Military Academy (RMA), at Université de Liège (ULg), and at Ghent University (UGent).

References

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 25, pp. 564–577, 2003.

[2] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proceesdings of the European Conference on Computer Vision*, vol. 4, pp. 279–290, 2004.

[3] J. Cheng and J. Yang, "Real-time infrared object tracking based on mean shift," in *Proceedings of the Iberoamerican Congress on Pattern Recognition*, pp. 45–52, 2004.

[4] A. Borghgraef and M. Acheroy, "Using optical flow for the detection of floating mines in IR image sequences," in *Proceedings of the SPIE Optics and Photonics in Security and Defence*, vol. 6395, Stockholm, Sweden, September 2006.

[5] A. Borghgraef, F. D. Lapierre, and M. Acheroy, "Motion segmentation for tracking small floating targets in IR video," in *Proceedings of the 3rd International Target and Background Modeling and Simulation Workshop (ITBMS '07)*, Toulouse, France, 2007.

[6] Z. Y. Wei, D. J. Lee, D. Jilk, and R. Schoenberger, "Motion projection for floating object detection," in *Proceedings of the 3rd International Symposium on Advances in Visual Computing*, vol. 4842 of *Lecture Notes in Computer Science*, pp. 152–161, 2007.

[7] O. Barnich and M. Van Droogenbroeck, "ViBe: a powerful random technique to estimate the background in video sequences," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 945–948, April 2009.

[8] P. M. Jodoin, J. Konrad, and V. Saligrama, "Modeling background activity for behavior subtraction," in *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '08)*, pp. 1–10, 2008.

[9] A. D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice-Hall, New York, NY, USA, 2002.

[10] M. Piccardi, "Background subtraction techniques: a review," *IEEE Transactions Systems, Man and Cybernetics*, pp. 3099–3104, 2004.

[11] S. Y. Elhabian, K. M. El-Sayed, and S. H. Ahmed, "Moving object detection in spatial domain using background removal techniques—state-of-art," in *Proceedings of the Recent Patents on Computer Science*, vol. 1, pp. 32–54, 2008.

[12] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, 2008.

[13] Z. Zivkovic., "Improved adaptive gausian mixture model for background subtraction," in *Proceedings of the International Conference on Pattern Recognition*, pp. 28–31, 2004.

[14] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[15] K. M. Ochi, *Ocean Waves*, Cambridge University Press, Cambridge, UK, 1998.

[16] W. P. Pierson and L. Moskowitz, "A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigordskii," *Journal of Geophysical Research*, vol. 69, pp. 5181–5190, 1964.

[17] K. Hasselmann, T. P. Barnett, E. Bouws, D. E. Carlson, P. Hasselmann, K. Eake, et al., "Measurements of wind-wave growth and swell decay during the joint north sea wave project (JONSWAP)," *Deutsche Hydrographische Zeitschrift*, vol. 8, no. 12, 1973.

[18] R. Stewart, *Introduction to Physical Oceanography*, Texas A & M University, 2008.

[19] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: principles and practice of background maintenance," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, vol. 1, pp. 255–261, 1999.

[20] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, 1997.

[21] C. Stauffer and E. Grimson, " Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 6, pp. 246–252, 1999.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, 2000.

[23] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.