# Waviz: Spectral Similarity for Object Detection

Christopher R. Wren and Fatih Porikli
Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA, 02139, USA

## Abstract

*Previous attempts to perform figure-ground segmentation have universally made the assumption that observations of the scene are independent in time. In the vocabulary of the stochastic systems literature: the individual pixels are taken to be samples from a stationary, white random processes with independent increments. Many scenes that could loosely be referred to as static often contain cyclostationary processes: meaning that there is significant structure in the correlations between observations across time. A tree swaying in the wind or a wave lapping on a beach is not just a collection of randomly shuffled appearances, but a physical system that has characteristic frequency responses associated with its dynamics. Our novel method leverages this fact to perform object detection based solely on the dynamics, rather than the appearance, of the pixels in a scene. Results are presented for a challenging scene containing wave activity in the background that visually masks a low-contrast foreground target.*

## 1 Introduction

The main contribution of this work is an algorithm, called *Wave Vision*, or Waviz, that explicitly harnesses the scene dynamics to improve segmentation. This allows Waviz to correctly interpret scenes that would confound appearance-based algorithms by having high-variance distractors in the presence of low-contrast targets, specifically when the distractors are well modeled as cyclostationary random processes. This is often the case, since real-world physics often induces near-periodic phenomenon in the environment: the motion of plants driven by wind, the action of waves on a beach, and the appearance of rotating objects. These periodic patterns are so strongly constrained by the relevant physics that efforts have been made to use then to recover camera calibration parameters by observing them[10]. In this work we merely strive to capture and exploit these patterns for segmentation.

## 2 Background

There is a rich literature that addresses the problem of detecting objects of interest in a scene that is unified by the definition of interesting: something is interesting if it is sufficiently different from a model of the stationary scene viewed through a stationary camera. The simplest of these methods assume that the scene is truly static, so that the uninteresting variability in the scene is adequately described by a unimodal, zero-mean, white, Gaussian noise process [13]. More complex systems include mechanisms for rejecting lighting changes as uninteresting, such as variablility caused by cast shadows [3].

Finally, there is a class of algorithms that allow the scene to be non-static. These algorithms represent the background as a multi-modal process [8], where each mode is a static model plus a zero-mean, white, Gaussian noise process [12]. The literature is far too varied to review here, however all these techniques have at their core the common assumption of a white process, that the observation process has independent increments [11].

## 3 Cyclostationarity

The independent increments assumption means that two samples drawn from the same pixel location will be independent. They may be drawn from the same probability distribution, but they will be independent samples from that distribution. The goal of the segmentation algorithm is to decide if the samples are drawn from the background distribution, or from some other, more interesting distribution. By assuming independent increments, these algorithms are relying completely on the appearance of the scene. Let's examine the case of a tree blowing in the wind. The multi-modal background models[12] would build up separate modes to explain, say sky, leaf, and branch appearances. As the tree moves, the individual pixel may image any of these. The independent increments assumption says that these different appearances may manifest in any order. However, we know that the tree will move with a characteristic frequency response that is related to it's physical composition. That characteristic response places constraints on

the ways that the library of appearances may be shuffled.

Specifically, given two samples from the observation process: $x[k]$ and $x[l]$, the independent increments assumption states that the autocorrelation function $R_x[k,l]$ is zero when $k \neq l$:

$$R_x[k,l] \triangleq E\left[x[k]x^*[l]\right] \quad (1)$$
$$= \sigma^2\delta[k-l] \quad (2)$$

where $\sigma^2 = E\left[x[k]x^*[k]\right]$ is the sample covariance and $\delta[k-l]$ is the discrete-time impulse function. This is correct when the process is stationary and white: such as a static scene observed with white noise. For a situation where the observations are driven by some physical, dynamic process, we can expect that the dynamics will leave their spectral imprint on the observation covariance. So if the process is simply periodic, then we would expect to see very similar observations occur with a period of $T$ samples, so in contrast to the above model:

$$R_x[k,k+T] \neq 0$$

We say that this process is cyclostationary if the above relationship is true for all time. More generally, wide-sense cyclostationarity is defined as [11]:

$$\mu[k] = \mu[k+T]\forall t \quad (3)$$
$$K_x[k,l] = K_x[k+T,l+T]\forall k,l \quad (4)$$

where $K_x[k,l]$ is autocovariance function for processes that are not zero-mean. These types of processes can be more complex than the simply periodic, and are characterized by significant structure in their autocorrelation functions, as illustrated in the self-similarity matrix shown in Figure 1. Figure 2 shows the sample trace from a pixel that is observing lapping waves on a beach.

process is said to be harmonziable if its autocorrelation can be reduced to the form $R_x[k-l]$, that is, so that the autocorrelation is completely defined by the time difference between the samples. It is possible to estimate the spectral signature of harmonziable, cyclostationary processes in a compact, parametric representation utilizing the Fourier transform [2]. Figure 3 shows an example transform of the same pixel as Figures 2 and 1 In the case of the evenly sampled, discrete observation processes we encounter in computer vision, we can use the efficient Fast Fourier Transform (FFT).

In Waviz we propose to estimate a background model that captures these spectral signatures, and then use those signatures to detect changes in the scene that are inconsistent with these signatures. By leveraging these dynamic constraints, we should be able to achieve higher specificity than a segmentation algorithm that ignores these constraints. Below we show results that demonstrate the ability of this kind of model it find low-contrast targets embedded in high-variance, dynamic scenes that are largely inaccessible to classic techniques.

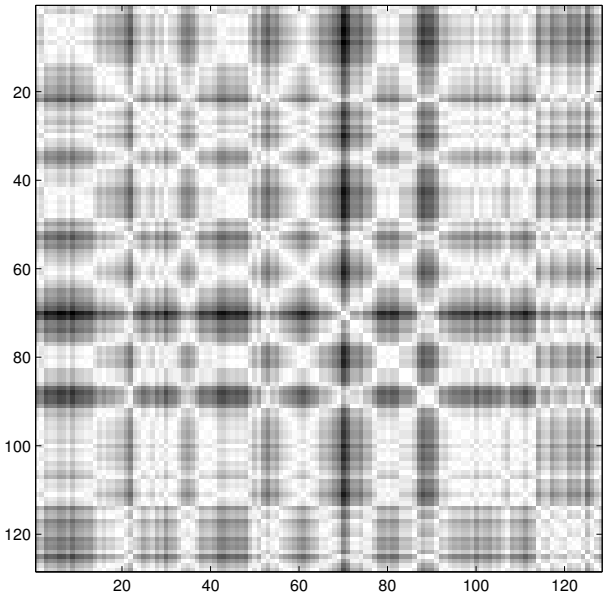


Figure 1: self-similarity matrix from the near-field water region at pixel$(320, 400)$

# 4 Using Spectral Similarity

Of course, the literature is not devoid of work that utilizes spectral fingerprints as a classification feature. However all the previous work on using spectral methods to classify activity have in common that they apply the spectral methods only to the foreground objects [6, 4, 1, 9, 7]. That is, objects that are either stationary in the frame, or have been extracted from the scene and stabilized by some other process, typically one of the segmentation schemes discussed above combined with some kind of tracker framework.

The literature on temporal textures contains some work on building searchable representations for video databases that would allow the system to recognize activity. These representations needed to be compact for storage in databases, and concise for quick indexing. As a result they involve summarizing the spectral content as a single number, for example, as the ratio of harmonic power to non-harmonic power in the signal. This involves explicitly attempting to extract features from the signal in the Fourier domain[6, 4]. We make no prior assumptions about what features will be interesting in the frequency domain, and instead use the Fourier signal directly.

The surveillance literature also contains work on spectral fingerprints that focuses instead on analysis of the full

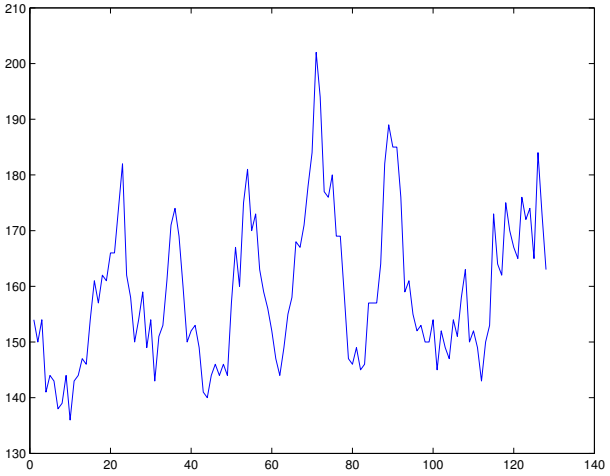


Figure 2: Sample trace from the near-field water region at pixel$(320, 400)$

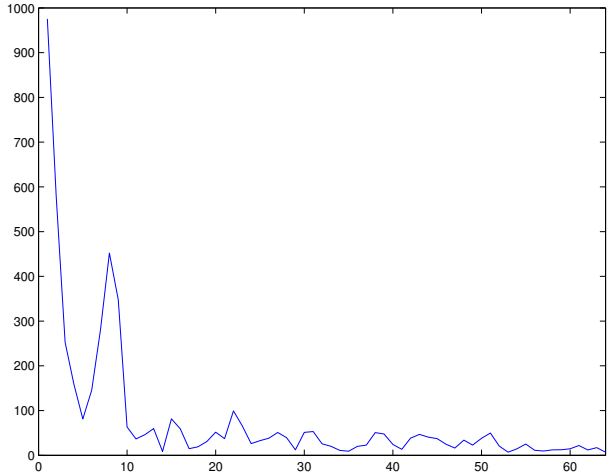


Figure 3: Fourier transform of the near-field water region at pixel$(320, 400)$

process autocorrelation function[1, 9]. This work is aimed at detecting pedestrians and pedestrians with sprung masses (backpacks, satchels, and the like). Howver, the word detect in somewhat misleading in this work. The method classifies targets into pedestrian, and non-pedestrian classes after they have already been extracted from the scene using classical techniques. As a result, these works, like the others above, make an independent increments assumption about the scene dynamics, even while exploiting rich descriptions of foreground object dynamics.

One system that is closely related to Waviz is the work by Liu and Sarkar[5] that uses *a priori* models of the periodicity in pedestrian motion to aid in the detection and segmentation of pedestrians in video. It is similar to this work in that it uses models of periodicity to segment video. This is in contrast to the work mentioned above that uses periodicity to classify motion only after it is segmented. It is different from this work in that it is using engineered models of a particular foreground process that is deemed *a priori* to be interesting: pedestrian motion. Waviz instead builds models *in situ* of the observed scene. It is therefore sensitive to anything that is sufficiently different from that situation-specific scene. Both use periodicity for segmentation, but the Liu work has a pointedly narrow definition of interest, while Waviz adopts a very inclusive definition.

# 5 Implementation

We begin by accumulating sample sequences for each pixel from a number of frames of video. Each of these sequences serves as an example of the periodic behavior of a particular pixel in the image. An example sequence is shown in Figure 2. The sample shown in that figure represents a single pixel over 128 frames of video.

These sequences, $x[n]$ are used to initialize the background model for each pixel. We extract a harmonic series representation, $a[k]$ using a discrete Fourier transform, such that:

$$x[t] = \sum_{k=0}^{N-1} a[k] e^{i2\pi kt/N}$$

We use only the magnitudes $\|a[k]\|$ of the Fourier coefficients in the representation. We take this as an estimate of the spectral components in the autocorrelation function of the underlying scene process [2].

For each new sample $x[n']$ we extract a new harmonic series representation, $b[k]$ for the current observation sample using $x[n']$ and the window of $N-1$ previous samples. We take this to represent the process underlying the current observations.

To determine if these two samples sequences were generated by the same underlying process, we compute the $L_2$-norm of the difference between the two harmonic series:

$$d = \langle a, b \rangle = \left( \sum_{k=0}^{N-1} (\|a[k]\| - \|b[k]\|)^2 \right)^{\frac{1}{2}}$$

This distance, $d$ provides a measure of the difference between the underlying processes. Small distances are taken to mean that the samples are drawn from the same process, and therefore represent observations consistent with the scene.

The length of the window, $N$, is a parameter that must be chosen with care. If the window is too small, then low-frequency components will be poorly modeled. However large windows come at the cost of more computation and

more lag in the system. We compare results from 32-, 64-, and 128-point windows below.

# 6 Short-Time Fourier Transform

The short-time Fourier transform (STFT) is a Fourier-related transform used to determine the sinusoidal frequency and phase content of a signal as it changes over time. Simply described, a window function, which is non-zero for only a short period of time, is multiplied by the function to be transformed. The window functions are applied to avoid discontinuities at the beginning and the end of a set of data. The smaller these discontinuities are, the faster the side slopes drop. The window functions, such as a Gaussian, usually have a cone shape centered around zero. The data to be transformed is broken up into chunks, which usually overlap each other. Each chunk is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency. This can be written as:

$$S[k, \omega] = \sum_m x[n+m]w[m]e^{-j\omega m} \quad (5)$$

for signal $x[n]$ and window $w[n]$. One of the downfalls of the STFT is that it has a fixed resolution. The frequency resolution is set mainly by the size of the segment, although some benefits may be derived from using a higher count (zero-padded) FFT, especially when using small segment sizes. The segment size also determines the percent of the overall data stream processed in a single FFT. Thus the time resolution is also fixed by the segment size (and to a much lesser extent by the sharpness of the data tapering window).

The segment size thus controls the tradeoff between frequency resolution and time resolution. Choosing a wide window gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution. Optimizing the STFT usually involves (1) finding an appropriate segment size, (2) setting the density in time by adjusting the amount of redundancy or overlap between the segments, (3) zero-padding the FFT for small segment sizes to better render spectral maxima, and (4) choosing an appropriate data tapering window.

## 6.1 Adaptation

Since the signals we encounter are almost never truly stationary, we add a simple exponential update mechanism to the above algorithm. This consists of combining the current estimate of the harmonic sequence with the estimate of the scene's harmonic sequence:

$$a_{t+1}[k] = a_t[k] + \alpha(b_t[k] - a_t[k])$$

| | uni | multi | 16-pt | **32-pt** | 64-pt | 128-pt |
|---|---|---|---|---|---|---|
| 20 | 50% | 37% | 16% | **10%** | 12% | 32% |
| 15 | 52% | 38% | 16% | **13%** | 18% | 33% |
| 10 | 65% | 46% | 22% | **22%** | 28% | 42% |

Table 1: Error rates. Percent of missed detections for three levels of tolerance: 10, 15, and 20 pixels. See text for further explanation.

Where $\alpha$ is the exponential mixing factor that we set to $0.02$ in all our experiments.

# 7 Results

We tested our algorithm on 2000 frames of infrared video of a boat moving offshore in the presence of waves. The frames were taken from the `All-Zod2` sequence. Frames 600-2600 were chosen because the camera remains stationary during those frames. The boat is a low-contrast object in the infrared video, and the wave action is a significant, dynamic distraction for any background segmentation algorithm. We compared several versions of our algorithm: using the 16-, 32-, 64-, and 128-point FFT. We also tested unimodal and multi-modal background subtraction algorithms.

In an attempt to most directly demonstrate the performance of the background models, we present sequential discovery performance. That is, we assume that there is one target per frame, and in each frame we pick the most novel pixel location as our estimate of the location of the target. Figure 5 shows some examples of the images transformed into distance maps by the algorithm. Obviously this scheme could be improved on with any number of tracking algorithms, but we feel this most directly demonstrates the performance of the underlying scene models.

Figure 4 shows discovery results from three algorithms superimposed on an example frame of the sequence. The top row shows the independent increments methods (unimodal on the left, multi-modal on the right). The bottom row illustrates the much better performance of the adaptive FFT-based scene model using a 32-point (left) window and a 64-point window (right). the results are plotted on a sample image from the scene for context.

Table 1 gives some quantitative results for five algorithms: the unimodal model, a multi-modal model [8], and four FFT-based models of varying window sizes. The percentages indicate the number of frames where the algorithm's estimate of the boat's position was off by more than the threshold in the left column, in pixels. You can see that the 32-point, adaptive FFT performs the best of the five algorithms, with almost 4 out of 5 detections falling within 10 pixels of the target.

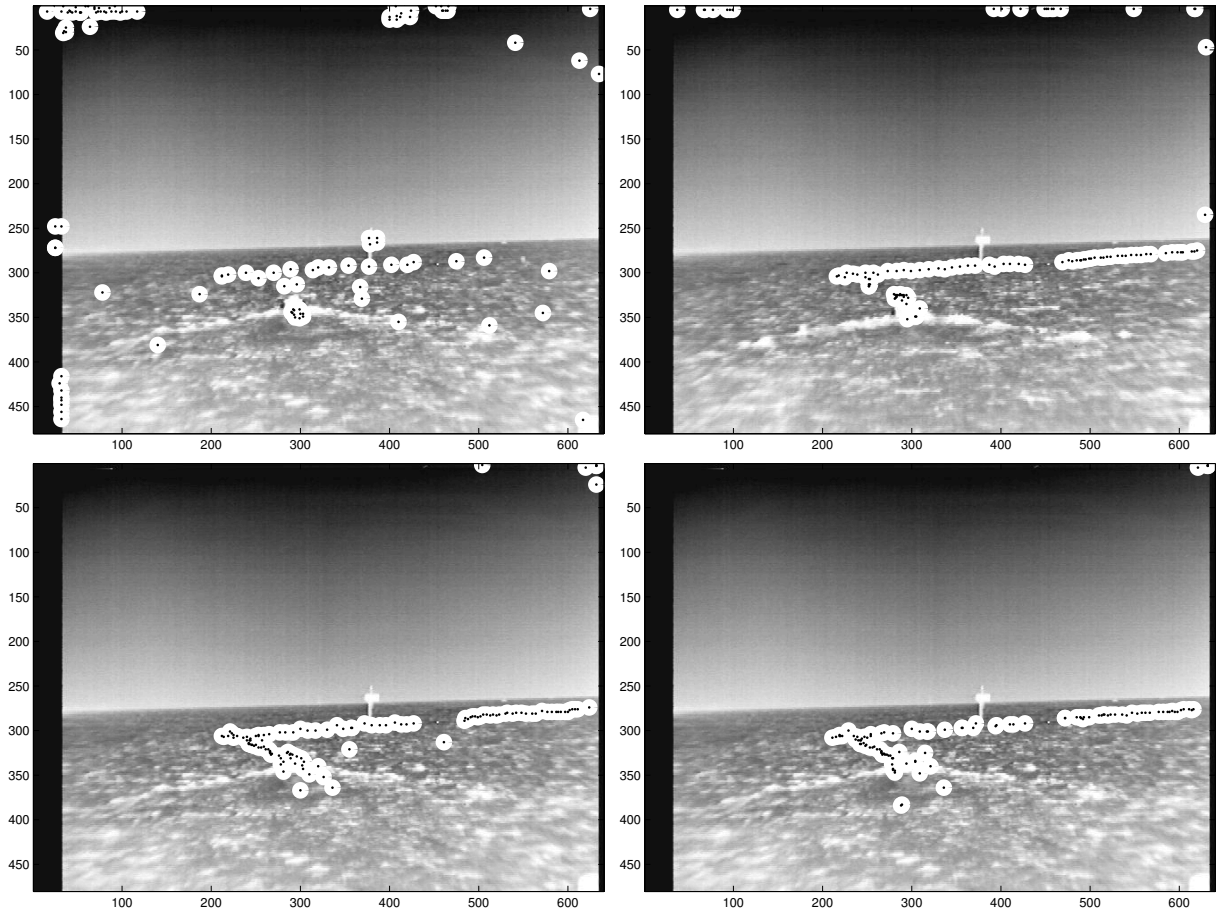

Figure 4: Tracking results. **Top:** unimodal (left) and multi-modal (right) background subtraction; **Bottom:** 32-point (left) and 64-point (right) adaptive FFT.

| uni | multi | 16-pt | **32-pt** | 64-pt | 128-pt |
|---|---|---|---|---|---|
| 280 | 161 | 98 | **92** | 102 | 99 |

Table 2: The root mean squared error between tracking results and ground truth.

Table 2 shows overall root mean squared error for the algorithms on the test sequence. Again the 32-point adaptive FFT algorithm is the best performer. However the 128-point adaptive FFT looks much better in this analysis: it significantly out-performs both of the Gaussian models. Given that the hit-rate performance if the 128-point algorithm was similar to the Gaussian mixture performance, this may indicate that the 128-point FFT is actually finding the target more often, but is reporting a highly biased position estimate that is causing near-hits to be often labeled as misses.

This assertion is given credence by the more detailed analysis of the results in Figure 2. We can see that the multi-Gaussian algorithm is very precise: if it find s the target it then it reliably gets the position correct to within several pixels. This is indicated by the sharp knee in the "multi" curve of Figure 6. The FFT-based algorithms seem to introduce a bias that corrupts the position estimate despite what is essentially a successful detection. We suspect that this is doe to lags introduced by the windowing of the sliding FFT. Eliminating that lag should be possible, but is deferred to future work.

# 8 Conclusion

We have presented a novel algorithm called Waviz that detects new objects based solely on the dynamics of the pixels in a scene, rather than their appearance. This is accomplished by directly estimating models of cyclostationary processes to explain the observed dynamics of the scene and then comparing new observations against those models. We have presented results that demonstrate the efficacy of this algorithm on challenging video.

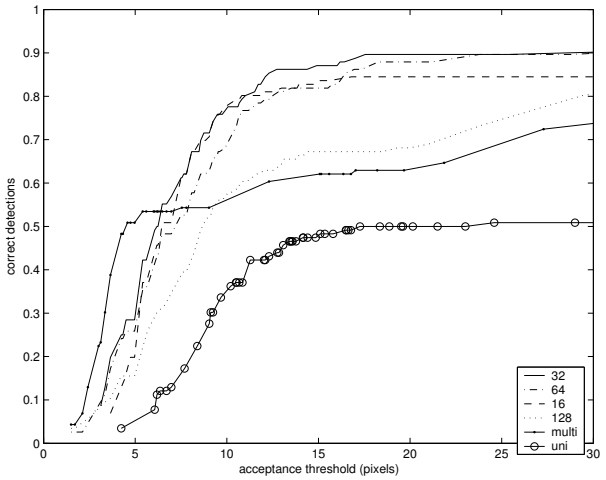


Figure 6: Analysis of hit ratio (vertical access) versus the threshold determining what is an acceptable hit.

## Acknowledgments

We would like to thank Professor Terry Boult and the University of Colorado at Colorado Springs for supplying the stimulating video sequences that were used in this paper.

## References

[1] Ross Cutler and Larry S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.

[2] Dominique Dahay and H. L. Hurd. Representation and estimation for periodically and almost periodically correlated random processes. In W.A. Gardner, editor, *Cyclostationarity in Communications and Signal Processing*. IEEE Press, 1993. http://citeseer.ist.psu.edu/33031.html.

[3] Ismail Haritaoglu, David Harwood, and Larry S. Davis. $w^4$: Who? when? where? what? In *Proceedings of FG'98*, Nara, Japan, April 1998. IEEE.

[4] Fang Liu and Rosalind W. Picard. Finding periodicity in space and time. In *International Conference on Computer Vision*. Narosa Publishing House, 1998. Also MIT Media Lab PerCom #435.

[5] Zongyi Liu and S. Sarkar. Challenges in segmentation of human forms in outdoor video. In *Perceptual Organization in Computer Vision*. IEEE, June 2004.

[6] Ramprasad Polana and Randal Nelson. Detecting activities. In *Computer Vision and Pattern Recognition*, New York, NY, June 1993. IEEE.

[7] Fatih Porikli and Tetsuji Haga. Event detection by eigenvector decomposition using object and frame features. In *Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, June 2004.

[8] Fatih Porikli and Oncel Tuzel. Human body tracking by adaptive background models and mean-shift analysis. In *Conference on Computer Vision Systems, Workshop on PETS*. IEEE, April 2003.

[9] Yang Ran, Isaac Weiss, Qinfen Zheng, and Larry S. Davis. An efficient and robust human classification algorithm using finite frequencies probing. In *Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, June 2004.

[10] Lisa Spencer and Mubarak Shah. Water video analysis. In *International Conference on Image Processing*, Singapore, October 2004. IEEE.

[11] Henry Stark and John W. Woods. *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice Hall, 2 edition, 1994.

[12] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition*, volume 2, Fort Collins, Colorado, June 1999.

[13] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

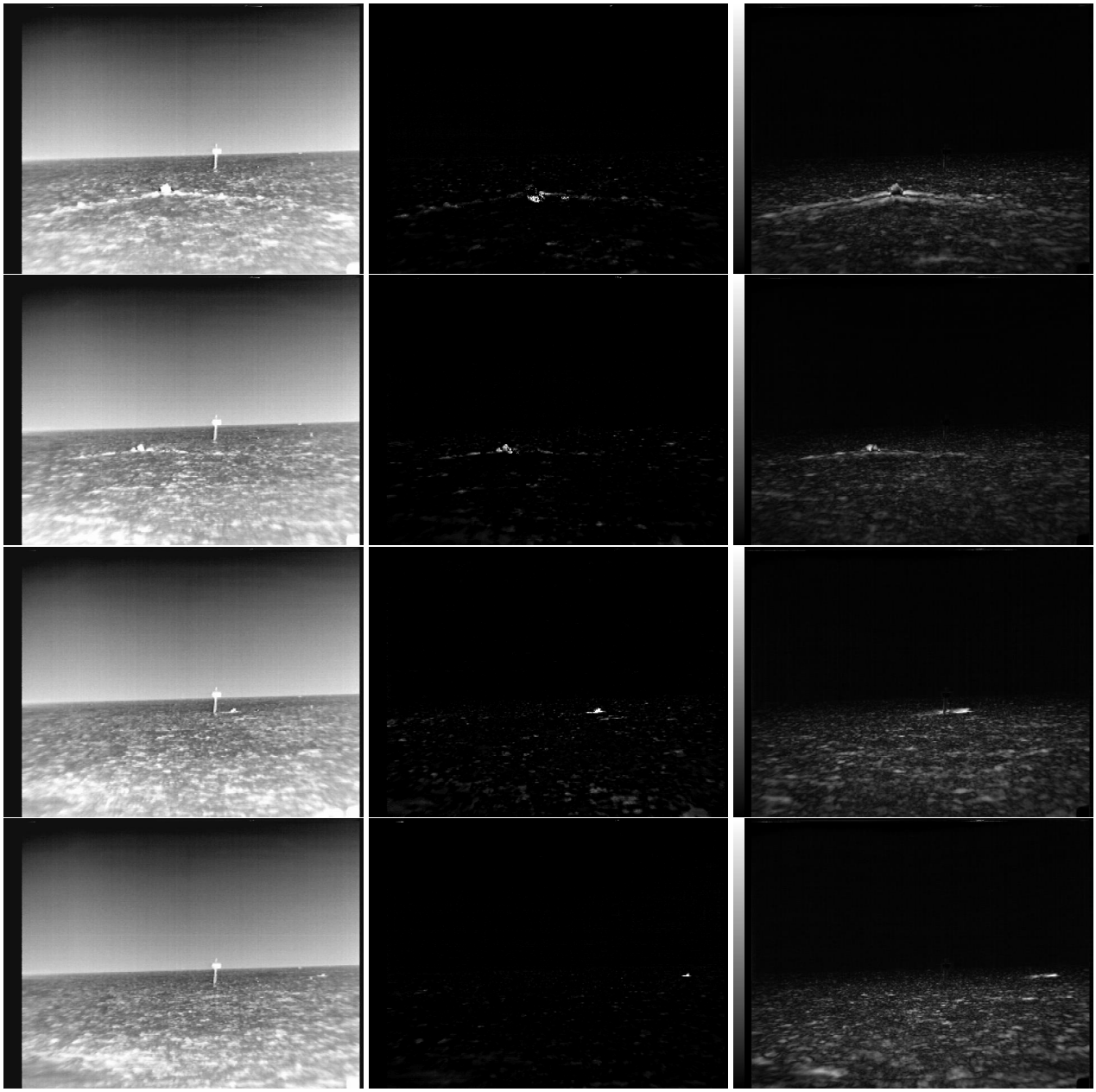

Figure 5: **Left:** Select frames from the sequence. **Center:** The distance transformed images that are generated by the multi-modal Gaussian background algorithm. **Right:** Corresponding images created by the 64-point FFT Waviz algorithm. Bright means novel. Frames are individually normalized.