Monitoring of Human Sport Activities at Sea

Rebeca Bastos
rebecaguerra@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisbon, Portugal

***Abstract*—The research problem addressed by this thesis concerns the people monitoring while performing human sport activities at sea, such as surf and swimming. The proposed solution uses video cameras installed on land, which are capturing real-time images of the sea, a problem that is largely unexplored. To analyze the videos captured by these cameras, automatic algorithms are used: saliency maps and probabilistic methods. Saliency maps methods are inspired in the human brain alertness mechanisms and are general in such way that they do not depend on the application neither on the images environment. On the other hand, the probabilistic methods use models capable of learning the image statistics. The different methods considered in this thesis are applied to real video sequences which were captured in Ribeira d'Ilhas beach, in Ericeira. The results of the automatic detection of people at sea are compared to the ones manually annotated and a statistical evaluation of system performance is provided. The proposed system, Saliency Residual Approach, achieves a false positive rate of 1.77, false positives per frame, for a detection rate of 90%.**

***Index Terms*—Maritime surveillance, Detection of People at Sea, Water Sports, Saliency Maps, Probabilistic methods**

I. Introduction

PORTUGAL has large maritime and coastal areas requiring important surveillance needs. Portugal has 1853 kilometers of coastal area: 950 kilometers in Mainland, 691 kilometers in Azores and 212 kilometers in Madeira, corresponding to 569 beaches in 2015. For security reasons, this area requires real time surveillance.

Vessel monitoring is achieved through a system named Automatic Identification System (AIS) that identifies each vessel up to 50 kilometers from the coast [1]. The system works as follows: each vessel equipped with (AIS), continuously sends information about the position, course and speed to the Coast Guard, in order to increase Maritime Domain Awareness (MDA) [1]. People surveillance along the coast is available only in a subset of beaches called supervised beaches, and it is performed by lifeguards. This means that there are extensive non-supervised areas where maritime activities are practiced without supervision, e.g. swimming, windsurfing, surfing and paddling. Thus, the existing way of people monitoring is limited either in effectiveness or in its duration.

One way to minimize the mentioned problems in people surveillance is by using real-time surveillance systems along the coastline, equipped with video cameras that are able to constantly monitor large maritime areas, as illustrated in Figure 1. These surveillance systems should be able to detect people at sea and activate an alarm or call the attention of a human operator when dangerous situations occur, *e.g.* when someone is far away from the coast. The solution of this surveillance problem is feasible with current technology but has not been explored yet.

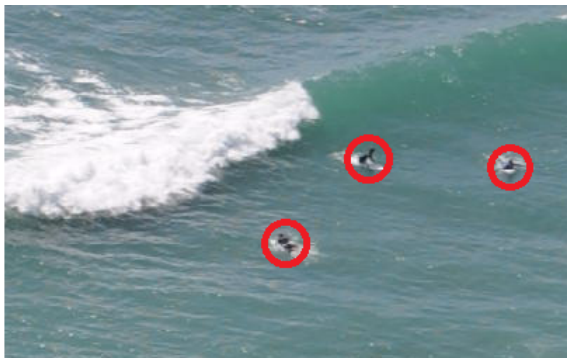

Fig. 1: Motivation scenario representing the goal of surveillance systems: identification of surfers or swimmers at sea (red circles).

As an alternative, would be the use of Unmanned Aerial Vehicles (UAVs) or drones equipped with video cameras for surveillance of both coastal and high sea regions. This type of video surveillance allows the detection of emergency situations, external threats, illegal migration or drug landing [2]. However, UAV are not suitable for solving the problem addressed in this thesis because the aircrafts are unable to be constantly flying and therefore do not guarantee continuously monitoring a certain area.

A. Problem Statement

The surveillance of human activities at sea is a new problem that has not been studied yet. This thesis proposes algorithms to detect people performing maritime activities such as swimming or surfing using cameras installed on land [3]. The lighting changes throughout the day and the changes in the dynamics on the background image (sea) hinders the identification of persons in maritime activities. In order to detect people at sea, this thesis adopts two types of methods: saliency maps and probabilistic methods. Saliency methods are faster and try to select informative regions in the image, allowing the use of more sophisticated methods in such regions, as it happens in the Human Visual System (HVS). These methods are general and do not depend on the application or the nature of image, which justifies its speed but also its limitations [4]. The probabilistic methods depend on statistical characteristics

of the image, they may be slower and may require a training phase to estimate the model parameters [5].

B. Thesis Contributions

This thesis addresses a novel problem: surveillance of human activities at sea. As main contributions this thesis applies different methods on real video sequences of people performing maritime activities and presents a statistical assessment of the algorithms' performance. It is also provided a dataset with people who are performing human sport activities manually annotated, over 300 frames.

C. Thesis Outline

Chapter II, describes several people detection algorithms. Chapter IV presents an experimental evaluation of the proposed algorithms, as well as detection results obtained with real data. Lastly, Chapter V and Chapter VI draws the main conclusions about this thesis and directions for future work.

II. People Detection Algorithms

While pedestrians detection is a well know problem, people detection at sea is a challenging and unexplored problem. People detection becomes much more difficult when people are located in the sea, due to the sea dynamic behavior. However, Humans have the ability to easily detect salient objects in a moving background. This ability motivates the study of saliency methods inspired in the HVS [6]. Another alternative, consists of using probabilistic models taking into account the statistical properties of the sea image.

Next sections presents four different methods: saliency maps, background subtraction, mixture of Gaussians and patch based models.

A. Saliency Maps

Saliency methods are fast methods that try to select informative regions in the image. These methods achieve good results in the identification of different classes objects, as demonstrated in [7]. Several saliency methods have been developed, *e.g.*, Saliency-Based Visual Attention [8], Graph-Based Visual Attention [9], the Spectral Residual Approach [10] and Adaptive Whitening Saliency [11], to name a few. Based on a recent comparative study [12] and based on a preliminary report we did [13], the chosen algorithm was the Spectral Residual Approach.

The Spectral Residual Approach method uses the fact that an image can be decomposed in two parts: prior knowledge, denoted as background, and innovation, corresponding to objects [10]. In order to determine the innovation part, objects, of an image I, the spectral residual approach is used: the difference between the natural logarithm of the amplitude spectrum of that image, $L(f)$, and a low-pass filtered version of it, $h(f) * L(f)$, where f is the frequency,

$$R(f) = L(f) - h(f) * L(f). \tag{1}$$

This method proceed as follows: First, it is computed the Fourier Transform,

$$\mathcal{F}[I(x)] = I(f) = |I(f)|e^{j\phi(I(f))}. \tag{2}$$

The residual spectrum is the difference between the logarithm of the amplitude spectrum, $L(f)$, and a low-pass filtered version of it, $A(f)$.

$$R(f) = L(f) - A(f) \tag{3}$$

where,

$$L(f) = ln|I(f)|, \tag{4}$$

and

$$A(f) = h(f) * L(f). \tag{5}$$

Lastly, the saliency map is constructed using the Inverse Fourier Transform:

$$S(x) = g(x) * \mathcal{F}^{-1}[e^{R(f)+j\phi(I(f))}]^2, \tag{6}$$

where $g(x)$ is a gaussian filter.

B. Background Subtraction

Background Subtraction methods are popular techniques used for pedestrian detection in urban areas [14]. The basic principle of background subtraction is to compare a static background frame with the current frame of a video sequence in order to detect regions where a significant difference occurs. This method distinguishes moving objects, denoted as *Foreground*, from static or slowly moving objects, denoted as *Background*.

To determine the background model for each pixel, several approaches are possible. One approach consists of using a set of values taken in the past at the same location, in order to create a background pixel [14]. Instead of using the same pixel in the past, other approaches may consists of using small neighborhood pixels, in the same frame, in order to create a background pixel.

The first approach involves learning the characteristics and create a background image, or frame, using a set of observed samples. First, the background pixel is computed using the mean value, given by:

$$\hat{I}_x = \frac{1}{N}\sum_{t=t_1}^{t_N} I_x(t), \tag{7}$$

where $I_x(t)$ is the intensity of pixel x at time t and N is the number of past samples. The second approach involves learning the model for a pixel x using pixels in its neighborhood, N_x. First, it is estimated the value of intensity for each pixel x, background pixel, using the mean value:

$$\hat{I}_x(t) = \frac{1}{k}\sum_{y(t)\in N_x} I_y(t), \tag{8}$$

where N_x is the considered neighborhood, I_y is the intensity of each neighbor pixel, y, and k is the number of neighbor pixels. In both approaches, the background pixel can also be computed using the median value:

$$\hat{I}_x = median\{\mathcal{I}_x\}, \tag{9}$$

where $\mathcal{I}_x$ is the list of intensities of pixel x over time, containing N past samples, or $\mathcal{I}_x(t)$ is the list of intensities in pixel x neighborhood, N_x, at time t.

In the test phase, the absolute difference between the background pixel, $\hat{I}_x$, and the test pixel, $I_x^*(t)$, is determined:

$$\Delta_x(t) = |\hat{I}_x - I_x^*(t)|. \tag{10}$$

The pixel x at time t is classified as Foreground if the absolute difference is greater than a specified threshold.

C. Mixture of Gaussians

Probabilistic methods consider the image as a realization of random signals characterized by a probability distribution. Some models try to represent the dependence between the image pixels in nearby locations while other methods assumes that each pixel is an independent random variable. In both cases, the model is learned from the data, *e.g.*, the mean and covariance, which is able to model the original data. In order to model the data different kind of distributions can be used, *e.g.*, Gaussian or Normal.

Given a set of observed samples, we want to estimate a probabilistic model based on a mixture of Gaussians:

$$p(I_x) = \sum_{k=1}^{K} w_k \mathcal{N}(I_x; \hat{\mu}_k, \hat{\sigma}_k^2), \tag{11}$$

where $p(I_x)$ is the probability density function for random variable I_x, K is the number of gaussians, w_k is the weigh of k-Gaussian and $\mathcal{N}(I_x; \hat{\mu}_k, \hat{\sigma}_k^2)$ is a normal distribution with mean $\hat{\mu}_k$, and variance $\hat{\sigma}_k^2$. Using the expression of the normal density, we obtain:

$$p(I_x) = \sum_{k=1}^{K} w_k \frac{1}{\sqrt{2\pi}\hat{\sigma}_k} e^{-\frac{(I_x - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2}}. \tag{12}$$

Consider a set of observations of a random variable, I_x, for each pixel x, and consider a set of parameters that have to be estimated $\Theta = (\hat{\mu}_k, \hat{\sigma}_k^2, w_k)$ where $k = 1, ...K$. In principle these parameters could be estimated using the Maximum Likelihood (ML) method. However the optimization problem cannot be analytically solved and there are no closed form expressions for the estimation. Therefore, the Expectation-Maximization (EM) method is used instead ML method. This algorithm is divided in two steps: Expectation and Maximization. The first step computes the expected values of the hidden labels (probability of each data point I_x being generated by each k-gaussian, $w_k(I_x)$). The second step updates the values of the different parameters:

$$\hat{\mu}_k = \frac{\sum_x w_k(I_x) I_x}{\sum_x w_k(I_x)}, \tag{13}$$

and

$$\hat{\sigma}_k^2 = \frac{\sum_x w_k(I_x)(I_x - \hat{\mu}_k)^2}{\sum_x w_k(I_x)}. \tag{14}$$

The iterations stops when the model parameters converge.

D. Non-Local Means

Denoising methods, as the name suggests, aim to improve the image quality and reduce the amount of noise. However, these methods could be a good approach to solve the current problem since the objects, in particular surfers, can be considered as noise in the image, at sea.

Non-Local Means is a powerful image denoising method whose goal is to recover the original image from a noisy measurement [15]. This method is based on patches comparison, thus it is searched for patches similar to the current one and it is computed a weighted average of those patches. This methodology assumes that an observed value I_x, in pixel x, is the sum of the "true" value, T_x, plus a noisy perturbation, NS_x, given by

$$I_x = T_x + NS_x \tag{15}$$

This method aims to estimate the intensity value of pixel x, $\hat{I}_x$, as a weighted average of patches centered in a general pixel y, N_y, in a small region of the image, S_x:

$$\hat{I}_x = \frac{\sum_{y \in S_x} w(x,y) I_y}{\sum_{y \in S_x} w(x,y)}, \tag{16}$$

where $\hat{I}_x$ is the estimated value for pixel x, $w(x,y)$ are the weights that depends on the similarity between the patch centered in pixel x and a patch centered in pixel y. Each weight is proportional to a similarity metric between the patch centered in pixel x, N_x, and the patch centered in pixel y, N_y, that depends on the difference between those patches intensities:

$$w(x,y) = e^{-\frac{\|I(N_x) - I(N_y)\|^2}{K\sigma^2}}, \tag{17}$$

where $I(N_x)$ and $I(N_y)$ are two vectors containing the intensity of all pixels in patches N_x and N_y, K is a normalizing constant and σ is the degree of filtering. Summarizing, patches similar to the current one receives large weights and different patches receives a small weight.

Since exploring the whole image is computationally demanding, we assume that y belongs to a search region (search window, S_x) centered at x. This choice influences the result since there may be similar patches outside this area. In this thesis we have slightly modified Non-Local Means algorithm. It was implemented a protection window around the current patch in order to prevent a search for similar patches in the same area of current one.

E. Discussion

This section presents a brief comparison between different methods described in Chapter II. Background subtraction (static background methodology) is a similar approach to the mixture of Gaussians if we assume that there is a single Gaussian per pixel. The main difference between both methods it is how a pixel is classified as a Foreground, meaning differences when is applied a threshold. While Background subtraction (static background) compares the absolute value of the error with a threshold:

$$|\hat{I}_x - I_x^*| > T_{BS(Static)}, \tag{18}$$

mixture models have a more complex way:

$$\frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(\hat{I}_x - I_x^*)^2}{2\hat{\sigma}^2}} < T_{MG(1G)}. \tag{19}$$

This expression can be simplified leading to:

$$(\hat{I}_x - I_x^*)^2 > -2\hat{\sigma}^2 ln(T_{MG(1G)}\sqrt{2\pi}\hat{\sigma}). \tag{20}$$

Lastly,

$$|\hat{I}_x - I_x^*| > \Big(-2\hat{\sigma}^2 ln(T_{MG(1G)}\sqrt{2\pi}\hat{\sigma})\Big)^{1/2}, \tag{21}$$

for

$$T_{MG(1G)}\sqrt{2\pi}\hat{\sigma} < 1. \tag{22}$$

From (18) and (21) it is possible to explain the main difference between background subtraction (static background) and mixture of Gaussians. Instead of doing a subtraction between the test pixel and the background pixel as in background subtraction method (static background), mixture models consider also the variance of pixel intensity over time.

III. Proposed System Architecture

The proposed system it operates in two different modes: Training and Detection (see Figure 2 and Figure 3). Training mode aims to estimate the model parameters, using experimental data (training data). The detection mode tries to detect human activities in new video data, using the model learned in the previous mode.

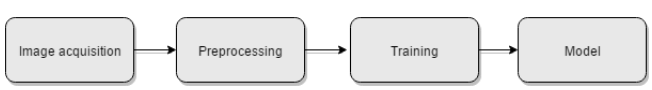

Fig. 2: System Architecture: Training mode.

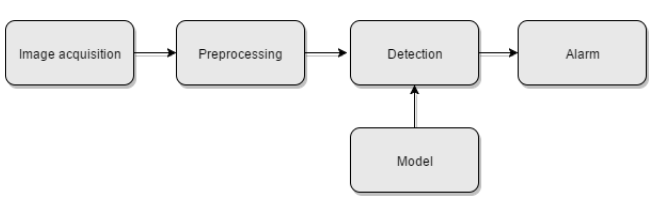

Fig. 3: System Architecture: Test mode.

The training mode starts by acquiring a set of video sequences (training data) that need to be labeled by a human. The camera is kept at a fixed position and the field of view is adjusted to cover the area of interest. The training sequences are then labeled by a human operator who defined the bounding boxes of all the objects of interest in the images. To keep the effort within reasoble bounds, this operation is done once every 24 frames (1 sec). The images are then preprocessed to convert them from the RGB format to gray scale format. The image resolution is also converted from 1270x720 pixels to a half. The model is then learned in the Training step. This step depends on the kind of method being used for the detection of swimmers.

In the test mode, the system received new video sequences and tries to detect all humans performing sport activities at the sea. The two first steps are the same except that the video data is not labeled by a human operator (except if it is used for the evaluation of the algorithm). Then, the video signal is processed by the detection algorithm, which takes into account the model learned during the training phase. The coordinates of the detected humans are then delivered to an alarm generator module which checks if there is any abnormal or dangerous situation.

IV. Results and Discussion

This chapter presents an experimental evaluation of the proposed algorithms using video recording of human activities at sea. First, we will present the video data used in this study. Then, we will define the evaluation criteria and present a statistical assessment of the proposed algorithms.

A. Video Data

We have considered four video sequences, each of them with a duration of 2 minutes and 30 seconds, meaning 3600 frames per video sequence and a frame rate of 24. Each frame have a resolution of 1270x720 pixels. Those sequences were acquired in Ribeira d'Ilhas beach, Ericeira, with a Nikon D5000 camera that was kept in a fixed position during the acquisition. The camera zoom was adjusted to be the sea the field of view. Figure 4 shows a Ribeira d'Ilhas map where were done the acquisitions, the camera position and the optical axis. The acquisition process were done in a sunny day and without wind.

Figure 5 shows an example frame extracted from those sequences. The video sequences are converted to gray scale images and processed without using the color information. This simplification is done in order to minimize the amount of memory required to process a video sequence. The image obtained after this conversion is also showed in Figure 6.

The sequences were manually annotated to locate the position of the swimmers and surfers. This procedure was done every 24 frames meaning that each video have 150 frames manually annotated with surfers bounding boxes. These annotations are denoted in this thesis as ground truth bounding boxes (BB_{Gt}). Figure 7 shows an example procedure.

B. Comparison with Ground-Truth

Given all ground truth bounding boxes (BB_{Gt}), and the regions that were classified as "surfers" by the algorithm,

Fig. 4: Ribeira d'Ilhas map illustrating where were done the video sequences acquisition, the camera position (orange rectangle) and the optical axis (orange line).

Fig. 5: Example image extracted from video sequences.

Fig. 6: Example image used in this thesis, gray scale image.

denoted as detected bounding boxes (BB_{Dt}), we need to find a methodology to match and compare both sets of bounding boxes. The matching between the two sets of bounding boxes is done according to the following rules [3]:

- Each BB_{Gt} may be matched at most once
- Each BB_{Dt} may be matched at most once
- BB_{Gt} and BB_{Dt} form a potential match if their overlap area is higher than zero

$$Overlap = \frac{area(BB_{Gt} \cap BB_{Dt})}{area(BB_{Gt} \cup BB_{Dt})} \tag{23}$$

- Detections with higher confidence, meaning higher overlap between BB_{Gt} and BB_{Dt}, are match first

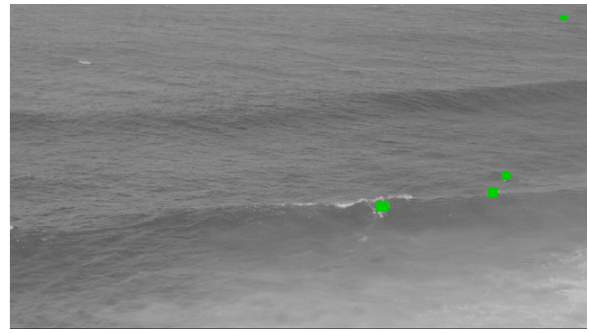

Fig. 7: Example image with ground truth bounding boxes (green boxes).

- If BB_{Dt} matches with several BB_{Gt}, we choose the match with highest overlap.

Figure 9 shows the result when applied the algorithm described above for the input image Figure 6. Figure 7 and Figure 8 show the BB_{Gt} (green boxes) and the BB_{Dt} (red boxes), respectively. Figure 9 illustrates the result of the matching between those two sets of bounding boxes, where yellow boxes is the matched BB_{Dt} with BB_{Gt}, meaning true positive detections and red boxes represents the unmatched BB_{Dt}, or false positives. If a BB_{Gt} unmatched with any BB_{Dt}, it is considered a false negative.

Fig. 8: Detected bounding boxes (red boxes).

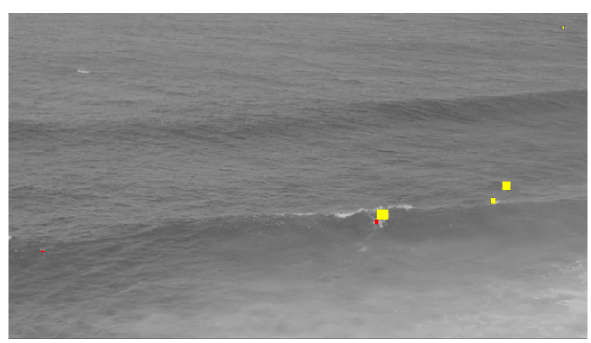

Fig. 9: Output image using matching methodology with ground truth: matched detected bounding boxes (yellow) and unmatched detected bounding boxes (red).

C. Performance Metrics

Several performance metrics can be used to evaluate the algorithms, *e.g.* Miss Rate curve, Recall-Precision curve, F-measure or ROC Curve [16], [17].

In this thesis, the chosen performance metrics are: Miss Rate curves [3], and Recall and Precision curves [18]. To plot it, the following statistical elements are used: true positive (TP), false positive (FP) and false negative (FN) or miss. In the context of monitoring maritime activities, TP is the number of correctly detected surfers, FP is the number of no surfers which were wrongly detected as surfers and FN is the number of surfers who were not detected. The variable TP is accepted as a surfer when there is an overlap greater than zero between the ground-truth and the area identified by the method. The following statistics are considered:

$$MissRate = \frac{FN}{TP + FN} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \tag{26}$$

D. Test Methodology

Four different methods are considered in this thesis: i) Spectral Residual Approach, ii) Background Subtraction (static background and neighbor pixels), iii) Mixture of Gaussians, iv) Non-Local Means. All these methods are estimated and evaluated using real data. In general, the operation of a decision-making system considers three phases:

- Training: This phase aims to estimate the model parameters, using experimental data
- Validation: This phase aims to optimize hyper-parameters (*e.g.*, number of Gaussians in the mixture) in order to improve the system performance
- Test evaluation: Performance the method with independent data

It should be stressed that only Mixture of Gaussians and Background Subtraction (Static background) require training phase.

Each phase should be performed using different videos. In this thesis the first two phases use the same video and the last phase uses a different video. Since the same video is used as training data and as validation data, these operations are done using K-fold cross validation [19]. In this methodology, the original video is partitioned into K subsets therefore requires K iterations. In each iteration, K-1 subsets are used as training data in order to create a model and the remainder subset is used to validate the model, validation phase. After K iterations, K results can be added to produce the final estimation. The advantage of K-fold cross validation is that each subset is used for validation exactly once and removes over-fitting problems. In this thesis, each validation video was divided into 10 subsets with 15 frames each.

E. Fine Tuning of Parameters

Each method depends on a set of hyper-parameters that should be chosen by the user. The validation phase aims to determine the value of each parameter in order to improve the algorithms' performance. The optimum values obtained by validation are shown in Table I.

TABLE I: Method Parameters and theirs used values for validation

Method	Parameter	Optimum value
Spectral Residual Approach	Size of filter window	3
BS (Static background)	Mean or Median	Mean
BS (Neighbor pixels)	Size neighbor window	2
Mixture of Gaussians	Number of gaussians	1
Non-Local Mean	Half length of patch	5
	Half length of search window	20
	Half length of protection window	12

F. Test Curves

The four algorithms will be assessed using two curves: Miss rate vs false positives per frame and Recall-precision curves, showed in Figure 10 and Figure 11, respectively.

The algorithm that achieves the best result is the *Spectral Residual Approach*. It is the algorithm that has, at the same time, lower values of both FP per image and miss rate (see Figure 10) and has higher values of both Recall and Precision (see Figure 11). The algorithm Mixture of Gaussians, Non-Local Means and Background subtraction (static background) revealed similar and the worst results. The algorithm Background subtraction (neighbor pixels) have revealed reasonable results.

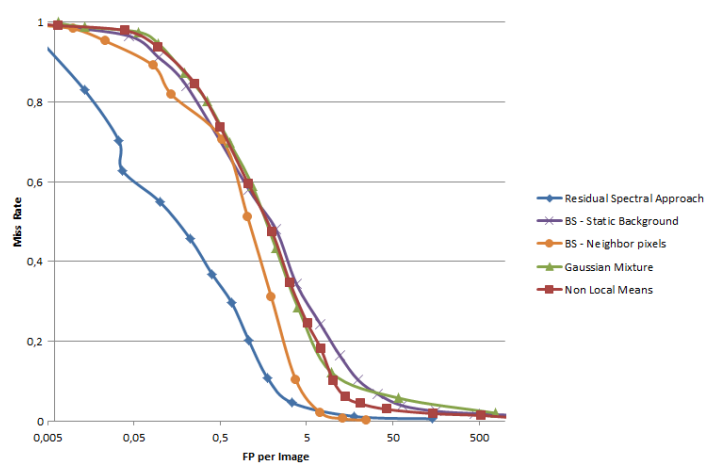

Fig. 10: Miss Rate experimental curves.

From Figure 10 and Figure 11 it is possible to conclude that methods without training phase give better results. However, since Mixture of Gaussians and Background Subtraction (static background) methods learns the background statistics, we expected better behavior.

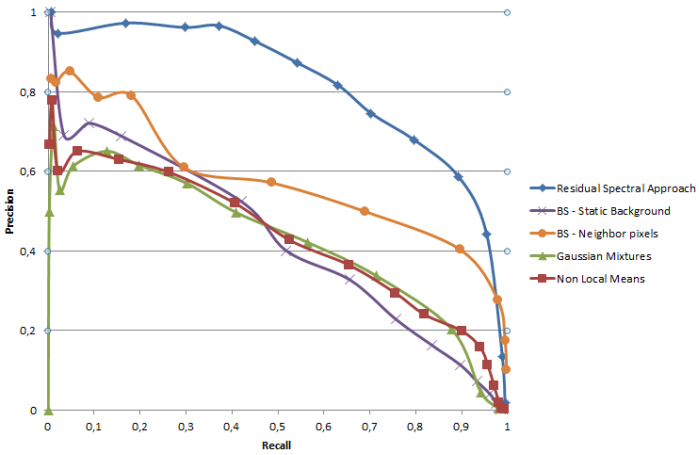

Fig. 11: Recall-Precision experimental curves.

G. Discussion

The previous section shows that Saliency Residual Approach gives better results in the video sequences used for testing, achieving 1.77 false positives per frame when miss rate, number of miss detections of all manually detection, is 10%. The background subtraction (static background and neighbor pixels), mixture of Gaussians and non-local means gives worse results than saliency residual approach, achieving 19.85, 3.74, 9.7 and 10.22 false positives per frame when miss rate is 10%. The obtained results are not good enough to solve the studied problem: people monitoring while performing human sport activities at sea.

As expected, background subtraction (static background) and mixture models with 1 gaussian reveal similar results with slight differences. Those differences could be explain by the comparison present in Section II-E.

V. Conclusions

In this thesis, we studied a new problem: the detection of people performing sport activities at sea using video cameras installed on land. This system have the potential to be very useful, in particular in countries with a large coastal area, since it aims to detect people at sea and activate an alarm or call the attention of a human operator when some dangerous situation occurs. The usage of fixed video cameras would be a good approach to this problem since the proposed system allows constantly monitoring large maritime areas.

Firstly, we presented different of methods: Spectral Residual Approach, Background Subtraction, Mixture of Gaussians and Non-Local Means. Those methods were chosen based on different studies. Secondly, we presented the experimental video sequences and a methodology to evaluate the performance of each algorithm. Lastly, we presented the experimental evaluation of the proposed algorithms, using video recording of human activities at sea. Spectral residual approach is the method that achieved better results, obtaining 1.77 false positives per frame when miss rate is 10% (detection rate of 90%). For example, in a frame with ten surfers, one of them is not detected on average and it is detected more than one false positive. Those results are not good enough and should be improved in the future before this kind of system can be considered as useful.

In general, we concluded that methods which requires training phase had worse results. This behavior is explained by the dynamic on the sea that occurs between the training phase and the test phase, in particular tide. The difference between the experimental results of background subtraction methods, using static background instead of neighbor pixels, 19.85 instead of 3.74 false positive for a miss rate of 10%, shows the weakness of modeling the sea, using a training phase. Experimentally, we find that mixture models, which needs a training phase to create a model, have worse results than we expected when identifying surfers or swimmers in the sea, 9.7 false positive for miss rate of 10%. This behavior could be explained also by the difficulty of modeling the sea. Non-local means performed 10.22 false positives for miss rate of 10%.

VI. Future Work

A negative aspect of the used approach is the amount of memory and computational time required to process a video sequence. In order to minimize the memory required should be used a real-time system instead of using a ”offline” methodology, meaning that all frames of video sequence are stored in memory. In order to minimize the computational time, an adaptive model should be used instead of a static model, in training phase. This model aims to minimize the training phase and include it in test phase, adapting the model when a change occurs, *e.g.*, the area where are breaking waves. The adaptive model should also improve the algorithms performance that need a training phase.

In this thesis, the experimental sequences were used in gray scale. It would be interesting to see the changes in methods performance when using color information. However, using the three color channels instead of using gray scale images is more computationally demanding.

References

[1] J. F. Vesecky, K. E. Laws, and J. D. Paduan, “Using HF surface wave radar and the ship Automatic Identification System (AIS) to monitor coastal vessels,” *IEEE International Geoscience and Remote Sensing Symposium*, pp. III–761–III–764, 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5417876

[2] A. Holst, B. Bjurling, and J. Ekman, “A Joint Statistical and Symbolic Anomaly Detection System: Increasing performance in maritime surveillance,” *IEEE Conference on Information Fusion*, pp. 1919–1926, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6290535

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: an evaluation of the state of the art.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–61, Apr. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21808091

[4] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, “Salient object detection and segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 409–416, 2014. [Online]. Available: http://mmcheng.net/salobj/

[5] S. Qi, C. Tao, C. Yang, and J. Tian, “A Robust Directional Saliency-Based Method for Infrared Small-Target Detection Under Various Complex Backgrounds,” *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 495–499, May 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6297996

[6] T. Kadir and M. Brady, "Saliency , Scale and Image Description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[7] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 73–80, 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5540226

[8] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *MIT Press*, pp. 545–552, 2006. [Online]. Available: http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2006_897.pdf

[10] X. Hou and L. Zhang, "Saliency Detection : A Spectral Residual Approach," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[11] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, Jan. 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0262885611001235

[12] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study." *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, Jan. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22868572

[13] R. Bastos, "Monitorização de Actividades Desportivas Marítimas," *Introdução à Investigação em Engenharia Electrotécnica e de Computadores - Instituto Superior Técnico*, 2015.

[14] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[15] H. Hu and Z. Xie, "A image denoising by modified nonlocal means method based on generalized cross-validation," *International Conference on Computational Problem-Solving*, vol. 2, no. 3, pp. 12–15, 2010.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[17] Y. Liu, E. Shriberg, and M. Park, "Comparing Evaluation Metrics for Sentence Boundary Detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. 185–188, 2007.

[18] K. H. Brodersen, C. S. Ong, K. E. Stephany, and J. M. Buhmann, "The binormal assumption on precision-recall curves," *International Conference on Pattern Recognition*, no. Section IV, pp. 4263–4266, 2010.

[19] M. Häfner, M. Liedlgruber, S. Maimone, A. Uhl, A. Vécsei, and F. Wrba, "Evaluation of cross-validation protocols for the classification of endoscopic images of colonic polyps," *IEEE Symposium on Computer-Based Medical Systems*, pp. 1–6, 2012.