# Automatic detection, tracking and counting of birds in marine video content

Roeland T'Jampens[1], Francisco Hernandez[1], Florian Vandecasteele[2] and Steven Verstockt[2]
[1] Flanders Marine Institute (VLIZ), Wandelaarkaai 7, 8400 Oostende, Belgium
[2] Data Science Lab, UGent-iMinds, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
e-mail: steven.verstockt@ugent.be

***Abstract*— Robust automatic detection of moving objects in a marine context is a multi-faceted problem due to the complexity of the observed scene. The dynamic nature of the sea caused by waves, boat wakes, and weather conditions poses huge challenges for the development of a stable background model. Moreover, camera motion, reflections, lightning and illumination changes may contribute to false detections. Dynamic background subtraction (DBGS) is widely considered as a solution to tackle this issue in the scope of vessel detection for maritime traffic analysis. In this paper, the DBGS techniques suggested for ships are investigated and optimized for the monitoring and tracking of birds in marine video content. In addition to background subtraction, foreground candidates are filtered by a classifier based on their feature descriptors in order to remove non-bird objects. Different types of classifiers have been evaluated and results on a ground truth labeled dataset of challenging video fragments show similar levels of precision and recall of about 95% for the best performing classifier. The remaining foreground items are counted and birds are tracked along the video sequence using spatio-temporal motion prediction. This allows marine scientists to study the presence and behavior of birds.**

***Keywords*— dynamic background subtraction, texture analysis, image classification, object detection, tracking, seabirds, marine environment.**

## I. Introduction

The Flanders Marine Institute (VLIZ)[1] has evolved into the central coordination and information platform for marine and coastal scientific research in Flanders, Belgium. The objective of VLIZ is to support and promote Flemish marine scientific research and to participate in local, national and international projects. The proposed research is linked to the European LifeWatch project[2] for monitoring biodiversity on earth. In the context of LifeWatch, VLIZ installed two PTZ cameras in a marine setting for visual surveillance. The first camera was placed at the Spuikom (Ostend), which is a local water mass attracting bird and people alike. The second camera is set upon the railing of a wind mill on the Thornton (sand)bank. The resulting video feeds, shown in Fig. 1, allow marine biologists to track the presence and behavior of birds on those sites. Manual operation of these camera systems, however, is not efficient due to fatigue, stress and the limited ability of human operators to perform this kind of tasks. To aid the scientists and avoid manual tallying, an automatic processing of the imagery is investigated.

[1]http://www.vliz.be/en/mission
[2]http://www.lifewatch.eu/

Automated object detection in a maritime and marine environment is a complex problem due to various factors that complicate the general video analysis approach. Camera motion, variety of objects and their appearance, highly dynamic background, meteorological circumstances, geographical locations and direction of the camera make the detection process challenging [1]. In order to deal with all these issues, an appropriate background model [2] is needed in combination with a classifier to discriminate between the objects of interest (i.e. birds) and other moving foreground objects. Furthermore, being able to track the objects of interest across consecutive video frames will facilitate detection, spatio-temporal behavior analysis and recognition. In this paper, we propose a methodology combining these three techniques and we particularly focus on the specific problem of bird detection in dynamic scenes. First of all, we build a background model to remove as much of the sea as possible, without removing any flying birds or birds resting on the water surface. Secondly, an image texture analysis is performed to classify the foreground candidates as water or bird. We investigate whether false detections (like water) can be eliminated, while maintaining the true detections (i.e. birds). This results in a number of validated foreground items per frame. Finally a spatio-temporal prediction technique is used to track these items along the video sequence.

The remainder of this paper is organized as follows. Related work in marine/maritime video analysis is discussed in Section 2. Subsequently, the three main building blocks of our algorithm are described in Section 3 to 5. Next, Section 6 presents our manually annotated VLIZ dataset (shown in Fig. 1), the evaluation process and results. Finally, Section 7 lists the conclusions and points out directions for future work.

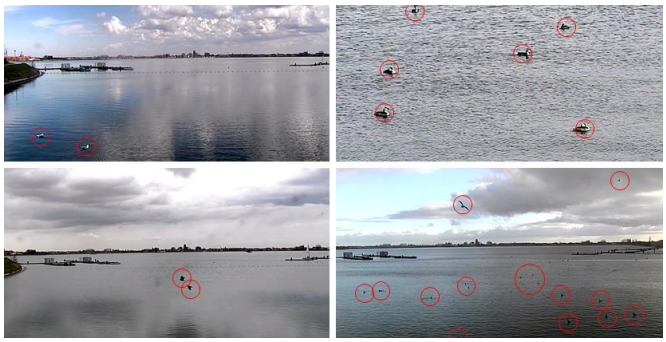

Fig. 1. VLIZ dataset with manually annotated bird labels.

978-1-4673-8910-5/16/$31.00 ©2016 IEEE

## II. Related work

The majority of video-based detection methods in a marine/maritime environment focus on vessel detection. Compared to birds, these objects are larger in size and have more robust/distinct features. Bloisi et al. [3], for example, use computational efficient Haar-like classification and spatio-temporal filtering to discriminate between vessels and other objects, e.g., reflections and wakes on the water surface. For bird detection, however, the discrimination step of this technique will fail. Furthermore, no real background modeling is used because the authors mention to have problems to cope with its dynamic behavior. For similar reasons, Rodrigo Da Silva Moreira et al. [4], mainly focus on maritime vehicles that arise above the horizon line. In order to detect the horizon line, Libe et al. [5], evaluated four different methods which tend to have a quasi-similar accuracy. However, only focusing on objects above the horizon line would limit the practical applicability of our approach and would complicate the tracking of birds.

The number of studies focusing on (sea) bird detection is still rather small, however, more recenlty research in this domain is beginning to appear more frequently. The skeleton based flying bird detection of Qunyu Xu and Xiaofeng Shi [6], for example, is based on the fact that the skeletal structure that most flying birds possess is rather similar and quite discriminative against other objects. SVM-based classification is used to label the simplified skeleton features with the appropriate object classes. The proposed method detects flying birds in side view with 90% accuracy, but needs improvement for adapting the wide variations in poses and viewpoints of free-flying birds. Histograms of Oriented Gradients (HOG) and Local Binary Pattern (LBP) are two other types of features that are commonly used to capture the edge, texture and local shape information of birds. The boosted HOG-LBP approach of Qing et al. [7], for example, detects 80% of seabirds correctly. False positives are mainly due to background problems. In order to tackle this issue, active area monitoring (for static backgrounds with a similar color distribution as the foreground objects) and dynamic background modeling can be used. The latter one is integrated in our approach. HOG features are also used in the method of Yoshihashi et al. [8]. This method serves as a basis for testing a novel bird dataset and is compared to a convolutional neural network (CNN) based method in more recent work [9], showing the effectiveness of rising CNN for general image classification tasks. CNN-based bird detection is also slightly investigated within this paper and will be the focus of future work.

In practical environments, such as our VLIZ context, birds tend to appear in low resolution even in a high resolution image, i.e. the monitoring system has to cover a wide field of view in order to perform a wide-range of tasks. It is important to take this into account when evaluating and comparing with other state-of-the-art approaches. A similar remark is made by Yoshihashi et al. [9]. The average size of the birds in their dataset is around 25 pixels, requiring recognition techniques that works on very low-resolution images.

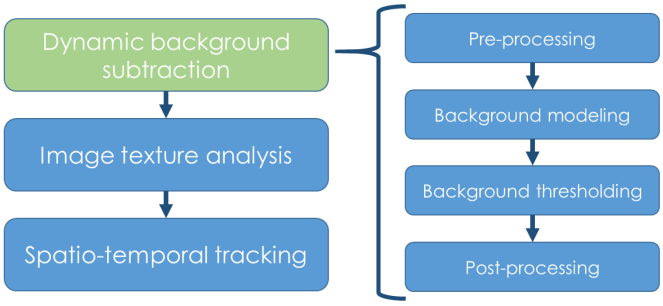


Fig. 2. Genaral overview of our dynamic background subtraction workflow.

The closest match to our solution is the system described in [1]. The author brings up many issues that arise in an intelligent video surveillance system for dynamic monitoring of objects in a maritime environment. Their system uses polynomial background estimation and classification for reduction of false detections. However, performance of the described detection and tracking algorithm may be limited in our bird-based scenario with low-resolution objects of interest. With a precision of 0.388 and recall of 0.516 their classifier based system is at the moment not nearly able of fully eliminating the false detections while remaining the true detections. However, the proposed future work in which additional features and a combination of classifiers for the sea- and air part will be examined looks promising.

## III. Dynamic background subtraction

Maritime environments represent one of the most challenging scenarios for background subtraction due to the complexity of the monitored scene [3]. Based on their performance reported in other works [10], two types of dynamic background subtraction (DBGS) techniques were selected and investigated using the workflow shown in Fig. 2. The first type defines individual pixels as fore- or background and is based on Temporal Median Filtering (TMF). The second type determines a block of pixels and is based on Local Binary Patterns (LBP) [11]. The latter block-based approach has a lower level of detail (pixel block) and is more complex, which makes it slower. As all investigated LBP techniques [12] did not perform well in our experiments, as shown in Fig. 3, these were excluded from further research.

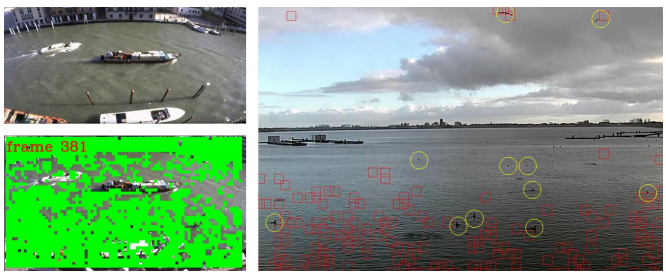


Fig. 3. False detection problem of Local Binary Patterns (LBP) based background subtraction: results on different types of maritime/marine video content show several detection problems for accurate bird counting/analysis.

Fig. 4. Results of the proposed dynamic background subtraction. Flying birds are more easily found than birds resting on the water surface. Different background model updating and thresholding approaches are investigated.

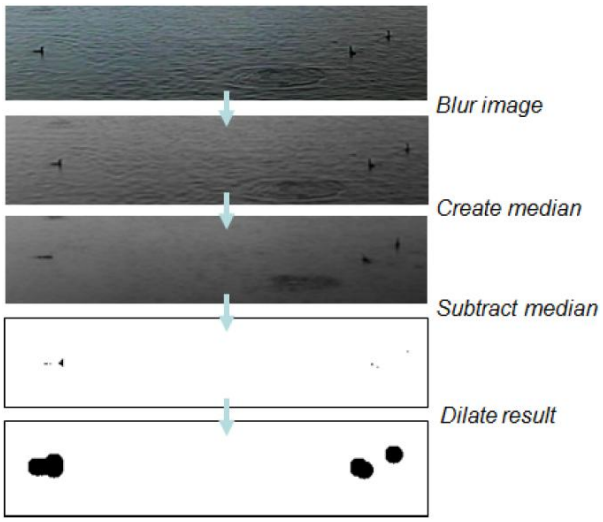


Fig. 5. Dilation improves the median filtering based foreground detection, making it possible to detect and analyze stationary objects such as birds resting on the water.

The pixel-based TMF approach keeps a limited set of video images in memory using a frame buffer. Based on this frame buffer we continuously update the background model with each incoming frame. For each pixel in the new frame, we determine the median value of the corresponding pixels in the frame buffer in order to perform a threshold-based background/foreground detection.

Three different approaches for updating the buffer are investigated. By default, the buffer is a sliding window in which the oldest frame is deleted first, i.e. a first-in-first-out strategy. An alternative approach uses a memoryless buffer in which frames are replaced at random, i.e. a Random Temporal Median Filter (RTMF) [13]. In this manner the background model covers a longer period of time/motion without enlarging the buffer. The technique is called memoryless because there is no link between the buffer index and the moment in time the frame was added to the buffer. The third option that we investigated is recursive, as only one frame is kept in memory and adjusted to each new frame. Approximate Temporal Median Filtering (ATMF) [14] saves the first frame received. If the new pixel value is higher than the model, the model is incremented by one. If a new pixel value is lower, the model pixel is decreased by one. For each of the investigated TMF approaches, all new frames are thresholded with the background model, resulting in the foreground mask. Pixels which differ more than the experimentally defined optimal thresholds are seen as foreground. Again, three different approaches were investigated. A first type of threshold uses a static threshold. However, as video content changes frequently (e.g. due to lighting and waves), the threshold should change accordingly. An alternative is using a relative threshold, in which the margin is expressed as a percentage instead of an actual pixel value difference. The last and the best approach is using a normalized threshold based on the mean and standard deviations of all pixel differences.

Median filtering is very accurate for detecting moving objects (see Fig. 4). However, for stationary objects, such as birds resting on the water, pixel values can be within the threshold of the median value. As a result, the foreground mask might not fully encompass all foreground objects. This is solved by dilating the foreground mask. In this way, stationary objects can be detected as foreground, as shown in Fig. 5.

## IV. Bird/water classification

To detect as many birds as possible, i.e. to limit the number of false negatives, some false positives (such as waves) must be allowed in the dynamic background subtraction. In order to filter out these false positives, we propose a classification mechanism based on image texture analysis (Fig. 6). Key points and corresponding gradient features are extracted from each detection result and are transformed into lower dimensional code words using the k-means clustering results of our training samples. Next, these code words are classified as 'water' or 'bird' by a linear SVM classifier (Support Vector Machines). During an offline training phase, labeled (bird, water) code words are fed to the classifier.

SIFT, i.e. the Scale-Invariant Feature Transform method of Lowe [15], is used as our baseline feature detector to which other approaches will be compared. SIFT describes an image by its most representative local characteristics and can be used for image stitching, object detection and object classification. The SIFT keypoints, i.e. circular image regions with an orientation, are represented by 128-dimensional vectors where the fastest varying dimension is the orientation. Some examples of VLIZ images with SIFT keypoints and orientations are shown in Fig. 7. Important to mention is that this technique is scale and rotation invariant in the 2D picture plane and to some degree to rotations in 3D.

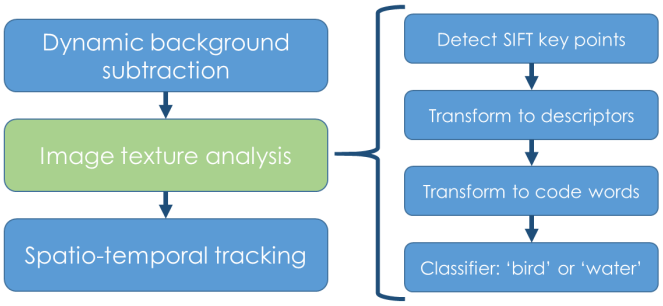


Fig. 6. General overview of our image texture analysis workflow.

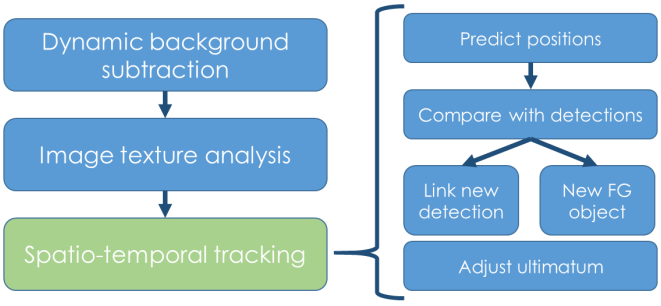


Fig. 8. General overview of our spatio-temporal tracking workflow.

For a classifier to work properly, all input data should have the same size and be sufficiently small to attain high performance. As our detected regions vary in size, the number of keypoints will greatly differ. Furthermore, a SIFT feature vector contains 128 bins, which is also rather big. Both issues complicate the training/classification process. However, features can be transformed into a combination of a limited number of visual code words [16]. Features of our training set are clustered with k-means, creating a small vocabulary of visual code words. The classifier is then trained on pictures described as a combination of these code words. For our baseline SIFT feature detector, a nonlinear SVM classifier with histogram intersection kernel [17] and only five code words is found to achieve the highest accuracy with a precision over 90% (as discussed in Section VI - Table 1).

Alternative feature extraction approaches based on SURF and HOG features have also been investigated. As shown in Section VI - Table 2, HOG features [18] seem to perform better than SIFT for the bird classification task. HOG is a dense feature extraction method for images that extracts features for all locations in the image as opposed to only the local neighborhood of keypoints like SIFT. Since the detected regions in our set-up can be very small, the number of SIFT features can be too low to discriminate between birds and water. Contrarily, the HOG descriptor technique counts occurrences of gradient orientation in localized portions, i.e. cells of 8x8 pixels, over the entire region of interset, leading to a higher accuracy. Furthermore, it is also important to mention that we don't use k-means clustering in the HOG based approach, which drastically decreases the computational cost (up to 80%) compared to our baseline SIFT-based approach.

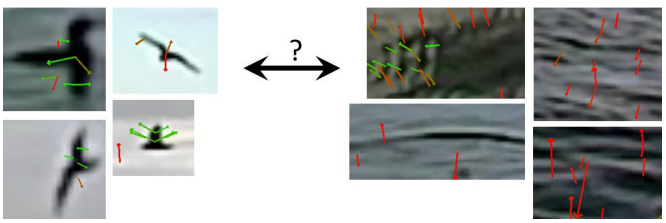


Fig. 7. Gradient analysis of bird and water examples from the VLIZ dataset.

## V. Spatio-temporal bird tracking

In order to track the detected birds across the video frames, individual detections must be linked to detections in previous frames. For each registered object, the new location is predicted on its last position and movement. If the bounding box of a newly found detection intersects with this prediction, the new detection is assumed the same object and the trajectory of the object is extended with the path to the new detection (shown in Fig. 9). Detections with no corresponding objects are seen as a new foreground object. Objects which have not been detected in a series of consecutive frames are deleted.

Before removal, object tracking information is written to a JSON-structured output file, as shown in Fig. 10. For each object, we log its first and last appearance and the spatio-temporal information of all its observations. These kind of loggings facilitate the querying of the video data and allows direct access to objects of interest, supporting VLIZ scientists in their study of the presence and behavior of birds.

## VI. Evaluation process and results

The proposed methodology is objectively evaluated using a ground truth labeled dataset of five videos coming from the VLIZ set-up, containing data recorded in different scenarios with varying light and weather conditions (as shown in Fig. 1). Different dictionary sizes, feature extraction methods and SVM kernels have been investigated.

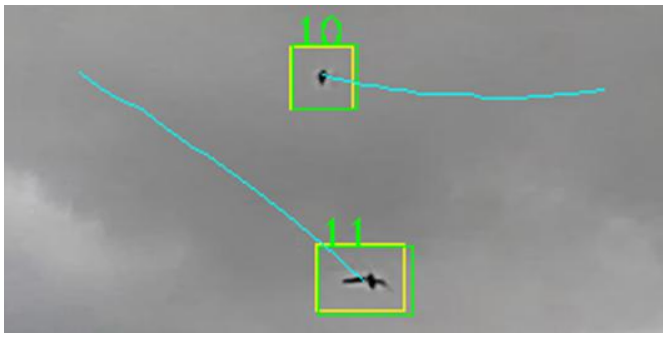

Fig. 9. Results of our spatio-temporal tracking of flying birds.

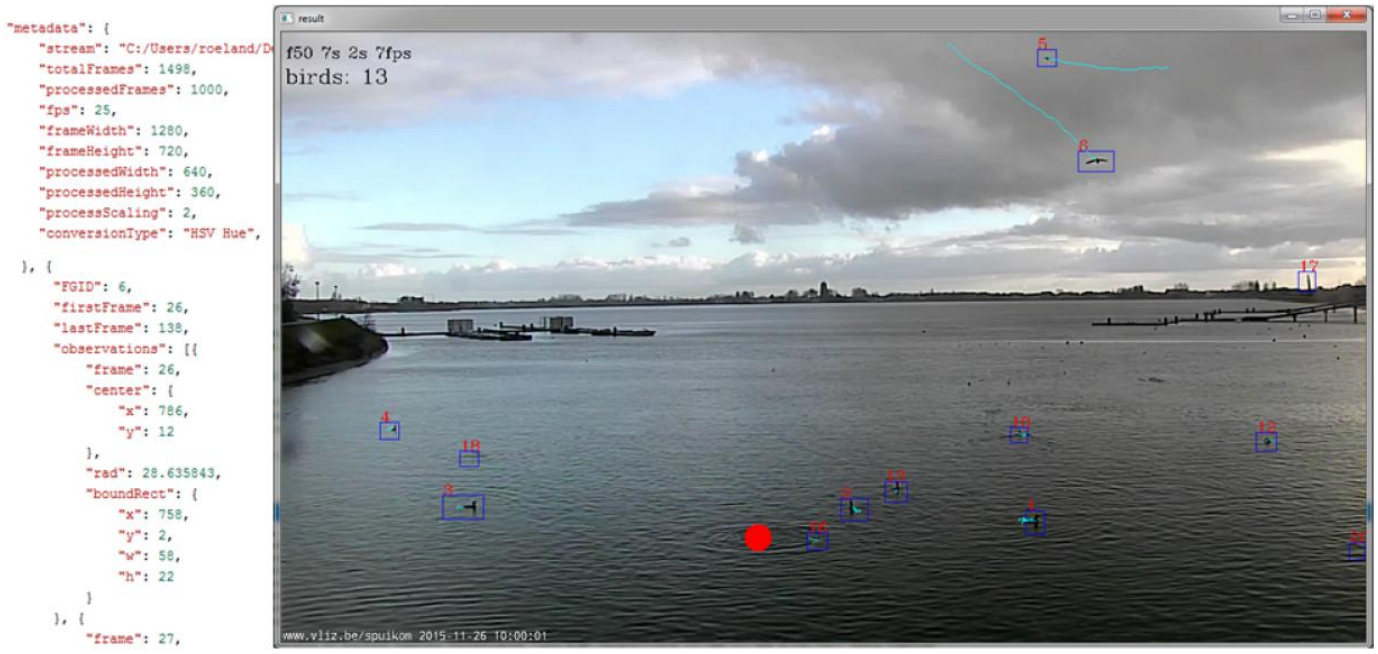


Fig. 10. Output of the proposed algorithm: JSON output file and annotated video stream.

Table 1 lists the results for our baseline SIFT based approach with three types of SVM kernels (linear, intersect and RBF) and a dictionary size of 5 and 10. A balanced training set of approx. 600 bird/water images and a test set of 10000 bird/water images were used in the evaluation process. In general, we achieve a high precision and recall. The best configuration consists of a dictionary size of 5 and an intersection kernel, resulting in 92% and 90% bird and water accuracy. Performance results of the classification step show that this configuration is also the fastest. We also have studied the impact of the type of median filter for dynamic background subtraction. No particular median filter can be seen as overall best, but a normalized threshold is clearly the better option in terms of accuracy and performance.

Precision, recall and F1-scores of alternative approaches for our baseline SIFT based approach are presented in Table 2. In this test, grid-search and shuffle split cross validation were used to avoid overfitting and to get the best parameters. A set of 3065 bird samples and 3483 water samples were used in this evaluation. In addition to SIFT, SURF and HOG have been tested to because both are mentioned quite often in literature. Again, different types of SVM kernels were evaluated. HOG features in combination with an SVM RBF kernel perform best with an average precision, recall and F1-score of 0.95. Important to mention is that no k-means clustering is used in HOG based classification, reducing the computational cost with 80% compared to the SIFT-based approach and making it more suitable for real time monitoring.

The confusion matrix of the best performing HOG - RBF approach contains 2892 true positives, 173 false negatives, 150 false positives and 3333 true negatives. In order to further decrease the number of false positives/negatives, we will investigate to incorporate temporal tracking information in the classification process. Furthermore, first tests have already been performed on state-of-the-art CNN architectures [19] for bird/water classification. The gaining importance of CNN for object detection and recognition will also be part of future work. Preliminary results of the Pyfaster object detection and recognition (shown in Fig. 11) with the COCO dataset (http://mscoco.org) show the feasibility of this approach.

The proposed detection results can be computed in quasi real-time and with low memory requirements when number of objects is low. However, the performance diminishes greatly if a lot of detections need to be classified. Redesigning the feature selection process in the bird/water classification, i.e. our computational bottleneck, will be further investigated.

Table 1. Accuracy and performance results for our baseline SIFT algorithm.

| Dictionary | Kernel | Bird | Water | Time |
|---|---|---|---|---|
| 5 | Linear | 92 % | 86 % | 91 s |
| **5** | **Intersect** | **92 %** | **90 %** | **89 s** |
| 5 | RBF | 90 % | 89 % | 95 s |
| 10 | Linear | 89 % | 89 % | 108 s |
| 10 | Intersect | 91 % | 91 % | 107 s |
| 10 | RBF | 91 % | 92 % | 107 s |

Table 2. Precision, recall and F1-score for different feature extractors.

| Feature | Kernel | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|
| | | *Bird* | *Water* | *Bird* | *Water* | *Bird* | *Water* |
| SURF | Linear | 0.86 | 0.89 | 0.88 | 0.87 | 0.87 | 0.88 |
| SIFT | Linear | 0.93 | 0.93 | 0.92 | 0.94 | 0.93 | 0.93 |
| SIFT | RBF | 0.93 | 0.93 | 0.92 | 0.94 | 0.93 | 0.93 |
| HOG | Linear | 0.93 | 0.93 | 0.91 | 0.94 | 0.92 | 0.93 |
| **HOG** | **RBF** | **0.95** | **0.95** | **0.94** | **0.96** | **0.95** | **0.95** |


Fig. 11. CNN Py-faster object detection with COCO dataset shows the feasibility of CNN based bird/water classification .

## VII. Conclusion

This paper proposes a methodology for automatic detection, tracking and counting of birds in marine video content and evaluates different state-of-the-art building blocks within this context. A dynamic background subtraction technique based on normalized median filtering gives fast and accurate results for the foreground object detection task. For bird/water classification of these foreground objects, the combination of HOG features and SVM RBF kernel performs best. With a best obtained precision and recall of 95% it is concluded that our current classifier based system is not yet able to fully eliminating the false detections while remaining the true detections. However, two suggestions for further improvement, i.e. temporal tracking information and CNN object recognition, are discussed within this paper and will be focus of future work. Finally, in order to track the birds across the video sequence, a bounding box based spatio-temporal tracker is proposed. This facilitates the querying of the video data and allows direct access to objects of interest, supporting scientists in their study of the presence and behavior of birds.

## Acknowledgment

The research activities as described in this paper were funded by the Flanders Marine Institute (VLIZ), LifeWatch (http://www.lifewatch.eu/), Ghent University, iMinds, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research-Flanders, the Belgian Federal Science Policy Office and the EU.

## References

[1] M. Hartemink. *Robust Automatic Object Detection in a Maritime Environment*. Master of Science Thesis, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology, Delft-the Netherlands, 2012.

[2] M. Piccardi Background subtraction techniques: a review In *IEEE International Conference on Systems, Man and Cybernetics*, proceedings, vol. 4, pp. 3099-3104, The Hague, Netherlands, October 2004.

[3] D. Bloisi, L. Iocchi, M. Fiorini and G. Graziano. Camera Based Target Recognition for Maritime Awareness. In *15th International Conference on Information Fusion (FUSION)*, proceedings, pp. 1982-1987, Singapore, July 2012.

[4] R. Da Silva Moreira, N. F. Favilla Ebecken, A. Soares Alves, F. Livernet, A. Campillo-Navetti. A survey on video detection and tracking of maritime vessels. *International Journal of Signal Processing*, 1, 47–60, 2016.

[5] T. Libe, E. Gershikov and S. Kosolapov. Comparison of Methods for Horizon Line Detection in Sea Images In *CONTENT 2012 : The Fourth International Conference on Creative Content Technologies*, proceedings, Nice-France, July 2012.

[6] Q. Xu and X. Shi. A Simplified Bird Skeleton based Flying Bird Detection In *11th World Congress on Intelligent Control and Automation*, proceedings, Shenyang-China, June-July 2014.

[7] C. Qing, P. Dickinson, S. Lawson and R. Freeman Automatic Testing Seabird Detection based on Boosted HOG-LBP Descriptors In *18th IEEE International Conference on Image Processing*, proceedings, Brussels-Belgium, September 2011.

[8] R. Yoshihashi, R. Kawakami, M. Iida and T. Naemura. Construction of a Bird Image Dataset for Ecological Investigations. In *22nd IEEE International Conference on Image Processing*, proceedings, Quebec-Canada, September 2015.

[9] R. Yoshihashi, R. Kawakami, M. Iida and T. Naemura Evaluation of Bird Detection using Time-lapse Images around a Wind Farm In *EWEA Wind Energy Event*, proceedings, Paris-France, November 2015.

[10] M. Cristani, M. Farenzena, D. Bloisi, and V. Murino Background subtraction for automated multisensor surveillance: A comprehensive review. *EURASIP Journal on Advanced Signal Processing* 1(24), 2010.

[11] M. Heikkila and M. Pietikainen A texture-based method for modeling the background and detecting moving objects *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657-662, 2006.

[12] J. Chen, G. Zhao and M. Pietikainen Unsupervised dynamic texture segmentation using local spatiotemporal descriptors In *International Conference on Pattern Recognition*, proceedings, pp. 1-4, Florida, USA, December 2008.

[13] O. Barnich and M. V. Droogenbroeck Vibe: A universal background subtraction algorithm for video sequences *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709-1724, 2011.

[14] S. S. Cheung and C. Kamath Robust techniques for background subtraction in urban traffic video In *SPIE - The International Society for Optical Engineering 5308*, proceedings, 2003.

[15] D. G. Lowe Object recognition from local scale-invariant features In *International Conference on Computer Vision*, proceedings, vol. 2, Corfu, Greece, September 1999.

[16] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray Visual categorization with bags of keypoints In *Workshop on Statistical Learning in Computer Vision*, ECCV, pp. 1-22, Prague, Czech Republic, May 2004.

[17] S. Maji, A. C. Berg, and J. Malik Classification using intersection kernel support vector machines is efficient In *IEEE Computer Vision and Pattern Recognition*, proceedings, Alaska, USA June 2008.

[18] N. Dalal and B. Triggs Histograms of oriented gradients for human detection In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Proceedings, pp. 886-893, San Diego, CA, USA, June 2005.

[19] S. Ren, K. He, R. Girshick and J. Sun Faster R-CNN: Towards real-time object detection with region proposal networks *Advances in neural information processing systems*, pp. 91-99, 2015.