Towards Maritime Videosurveillance Using 4K Videos

V. Marié[1,2(✉)], I. Bechar[1], and F. Bouchara[1]

[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, UMR 7020, Marseille, France
{vincent.marie,ikhlef.bechar,frederic.bouchara}@univ-tln.fr
[2] CS Systemes D'information, Paris, France
vicent.marie@c-s.fr

Abstract. This paper develops a novel approach to automatic maritime target recognition in the framework of near real-time maritime videosurveillance using super-resolved (i.e.; 4K) videos captured either with a static or with a moving video camera. The challenge of achieving a robust 4K video-based surveillance system is twofold. Firstly, the 4K video resolution (3840 × 2160 px.) constrains considerably the amount of video-processing for meeting the near real-time requirement. Secondly, maritime environment is very dynamic and highly unpredictable, thereby, rendering target extraction a difficult task. Therefore, the proposed approach attempts to leverage both temporal and spatial video information for achieving fast and accurate target extraction. In fact, since, the object rigidity assumption is implemented parsimoniously, i.e.; at key video locations, its real-time implementation, first, enables to quickly extract potential (sparse) target locations. Furthermore, we have shown, experimentally using many maritime videos, that maritime targets generally exhibit richer textural variations than dynamic background at different scales. Thus, secondly, a still image based multi-scale texture discrimination algorithm carried out around previously extracted key video locations allows to achieve final target extraction. An experimental study we have conducted both using our own maritime video datasets and publicly available video datasets have demonstrated the feasibility of the proposed approach.

Keywords: Maritime videosurveillance · 4K video
Spatiotemporal approach

1 Introduction

With the important recent advances in camera technologies and in computing resources, maritime videosurveillance has become an important research topic. The latter finds several real-world applications among which we can mention optimal monitoring of maritime traffic [3,9,21], seacoast security [18], prevention

Supported by the French Defense Agency (DGA), CIFRE no. 0004/2015/DGA.

A. Basu and S. Berretti (Eds.): ICSM 2018, LNCS 11010, pp. 123–133, 2018.
https://doi.org/10.1007/978-3-030-04375-9_11

of fraudulent maritime activities [7,16], situation awareness and prevention of asymetric threats (i.e.; a commercial or military ship being threatened by small maritime vehicles such as jetskis and inflatable boats). For instance, in the latter context, it is highly desired to be able to detect as early as possible small targets, hence, the need for highly performing computer vision hardware (e.g.; 4K video cameras) and software.

Although videosurveillance in controlled environments using conventional (e.g.; CIF) video formats is, now, a quite well understood computer vision topic, maritime videosurveillance still poses many challenges to the computer vision community. Indeed, traditionally, background subtraction algorithms [13,20,24,25] have achieved best state of the art performances in terms of detection accuracy and computational efficiency. Furthermore, in the goal of accounting for low video SNR, thereby, achieving more accurate object recognition, robust background subtraction algorithms–attempting to exploit small image neighborhoods, instead of single pixels–have been developed [6]. When the camera is moving and if the background is static, then a coupling of a background subtraction technique with a fast motion compensation algorithm, generally, yields very good results [23]. Nevertheless, the latter category of approaches is hardly applicable in maritime video-surveillance because of the dynamic background (e.g.; sea). Alternative approaches attempting to take advantage of spatiotemporal coherence of objects have been proposed [14,15]. More recently, deep learning based approaches have achieved astounding results in various application contexts, and in maritime vision in particular [4,8]. Nevertheless, our experiments have shown that the latter only achieve mitigated performances on maritime videos, above all, for detection of small and lowly contrasted targets. This can be partly explained by the fact that the latter category of techniques do not attempt to take advantage of the temporal video dimension in the goal of accounting for low maritime object contrast, and hence, for achieving better maritime object recognition performances.

In the contexts of maritime situation awareness and airborne maritime videosurveillance, every bit of a maritime scene is moving, including the camera, the background (i.e.; sea) and, obviously, maritime objects. Clearly, in such a context, none of the aforementioned techniques is suited. Consequently, novel computer vision algorithms using as little assumptions as possible are needed. Indeed, this work is part of a bigger research project aiming at developing state of the art maritime videosurveillance algorithms to automated situation awareness, and especially, for the fight against asymetric threats. Obviously, early detection of small ships requires near real-time processing of high definition video, typically 4K videos in our case. Therefore, in the remainder of this paper, we describe the approach that we have developed for maritime object extraction using 4K videos. The main contribution of this paper resides in the fact that video processing is carried out around key video locations while taking advantage both of the spatial and temporal video dimensions for achieving accurate and fast maritime target extraction.

2 General Method Overview

The main idea behind the proposed spatiotemporal approach to object recognition using 4K videos is motivated as follows. On the one hand, what mostly distinguishes a real-world object from a dynamic background is rigidity. On the other hand, we first hypothesize, then, we show experimentally using several maritime video datasets that maritime objects, generally, present richer textural features than maritime background. Furthermore, in the goal of achieving a near real-time 4K maritime videosurveillance system, we have efficiently implemented the latter idea with respect to the temporal and spatial dimensions of a video, respectively. Basically, a temporal algorithm allows to extract potential target keypoint locations as rigid ones using long-term keypoint tracking. Then, a spatial algorithm performs texture discrimination in the vicinity of the latter in the goal of achieving final target extraction. The workflow of the proposed approach is outlined in Fig. 1.

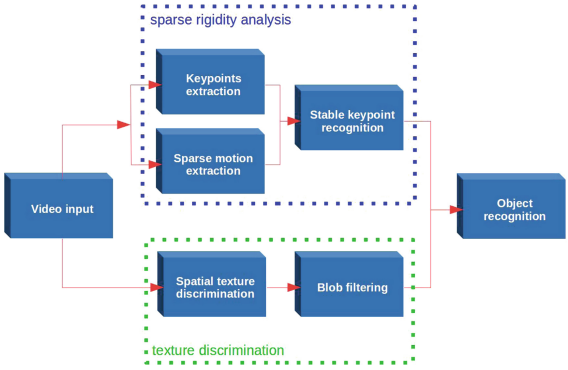


Fig. 1. Workflow of the proposed spatiotemporal approach.

3 Rigidity Analysis via Long-Term Keypoint Tracking

The key idea of exploiting the object rigidity hypothesis to achieve object extraction in video is not new. For instance, the authors in [22] have proposed a fundamental matrix based approach for performing multiple objects extraction in video. However, a major drawback of the fundamental matrix based type of approaches lays in their combinatorial nature, plus, in the difficulty of establishing enough "good" keypoint inter-frame correspondences. The latter issue is aggravated when the background is dynamic and/or the camera is moving which turns out to be, generally, the case in maritime videos. An alternative approach proposed in [2] attempts to estimate a dense rigidity criterion based on a timely $3D$ analysis of dense optical flow. But, the latter is hardly applicable in our case

due to unreliability of optical flow. Thus, we have proposed an alternative approach to rigidity analysis restricted to key video locations via keypoint tracking. Indeed, extraction of keypoints in individual frames, and establishment of their inter-frame correspondences is an easy, fast, and reliable process. Moreover, it makes sense to hypothesize that rigid object pixels tend to undergo less textural change in the course of time than dynamic background. Therefore, our approach to rigidity analysis using video consists in assessing the temporal textural variation at video keypoint locations via their long-term tracking as it will be in described in detail, hereafter.

3.1 Keypoint Extraction

The proposed approach to rigidity analysis in video begins with the extraction of key video locations. The latter are further tracked over time before they are declared as potential object keypoints, otherwise, permanently abandoned because they are, eventually, ranked as background. We have tested different existing keypoint extraction algorithms including SURF [1], ORB [19], and SIFT [11]. However, our experiments– using several maritime videos–have shown that SIFT outperforms considerably other existing keypoint extraction techniques both in terms of accuracy and repeatability, thus, we have finally opted for SIFT. The SIFT descriptor is, first, convolved with a $1D$ Gaussian profile ($\sigma = 2$), then, transformed into a probability distribution via mere division by a normalization constant, finally, stored in a discrete histogram of 128 bins.

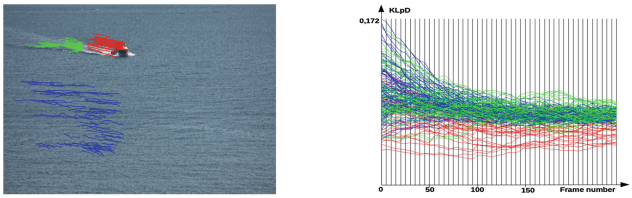
3.2 Keypoint Tracking

For the sake of computational efficiency, the extracted SIFT keypoints in the first video frame are tracked individually between frames using optical flow and the Kalman filter. The state vector of the Kalman filter consists here of the concatenation of the subpixel position (x_t, y_t) of a SIFT keypoint and its $2D$ velocity vector (u_t, v_t), in such a way that, the sate vector of the Kalman filter writes as $Y_t = (x_t, y_t, u_t, v_t)^T$, but, of which one may only observe a noisy version denoted by Z_t. Thus, the SIFT keypoint dynamic model writes as

$$\begin{cases} x_{t+1} = x_t + u_t + \epsilon_{t+1}^{(1)} \\ y_{t+1} = y_t + v_t + \epsilon_{t+1}^{(2)} \\ u_{t+1} = u_t + w_{t+1}^{(1)} \\ v_{t+1} = v_t + w_{t+1}^{(2)} \\ Z_{t+1} = Y_{t+1} + W_{t+1} \end{cases}$$

where it has been assumed that $\epsilon_{t+1}^{(1)}$, $\epsilon_{t+1}^{(2)}$, $^{(1)}_{t+1}$, and $w_{t+1}^{(2)}$ stand for four independent Gaussian random variables, W_T stands a 4-dimensional random Gaussian vector, last, subscript t denotes the time. Furthermore, we have used the method

developed in [5] for achieving robust optical flow estimation, and hence, accurate SIFT keypoint tracking. In short, such an optical flow estimation technique is based on the inversion of the Hessian matrix and which just happens to be well conditioned at SIFT keypoints. We refer the reader to [5,11] for more details.



(a) (b)

Fig. 2. Sparse rigidity estimation based on keypoint tracking. (a) Tracking of SIFT keypoints across 250 4K frames; (b) Temporal evolution of the proposed rigidity measure for different classes of SIFT keypoints. In blue: wave; in green: wake; in red: object. (Color figure online)

3.3 Keypoint Rigidity Analysis and Classification

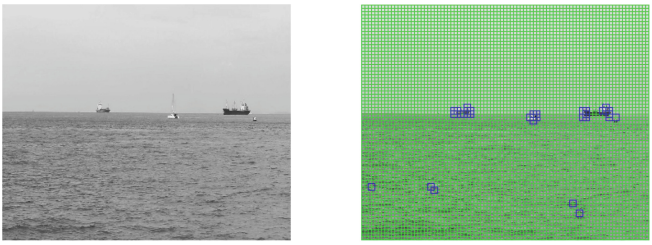
As aforementioned, SIFT keypoint classification as object *versus* background is based on the estimation of the long-term variation of the dynamic of the SIFT descriptor. Furthermore, since, we have modeled the latter in our approach as a probability distribution, it makes sense to use the Kullback-Leibler pseudo-Distance (KLpD) for deriving a measure of normalized SIFT keypoint textural variation. For obvious reasons, the latter is fairly correlated with any measure of rigidity, for rigid object regions are bound to undergo little textural change over time. Moreover, by basing a rigidity measure on a distance between probability distributions, one guarantees invariability against common geometric deformations of objects.

In fact, we have proposed a slight modification to the original KLpD for obtaining a symmetric and robust measure of textural variation over time, and which, overall, can account for the possible presence of near-zero values in a normalized SIFT descriptor, and slight shifts in the latter(mainly due to discretization). Thus, suppose two probability distributions P and Q, and define the divergence measure between P and Q as $D_{KL}(P||Q) = \sum_i D_i(P,Q)$ where one has $\forall i$:

$$D_i(P,Q) = \inf\left\{ \left|P(i)\log\frac{P(i)}{Q(i)}\right|, \left|Q(i)\log\frac{Q(i)}{P(i)}\right| \right\}$$

Robustness of $D_{KL}(P||Q)$ resides in that, given any couple of real positive numbers x and y, if $\inf\{x,y\} \to 0$, then $\inf\{|x\log\frac{x}{y}|, |y\log\frac{y}{x}|\} \to 0$, thereby, mitigating the contributions from too distinct couples of bins of P and Q,

respectively. Next, assume an arbitrary SIFT keypoint at time t, and denote by P_t the corresponding normalized SIFT descriptor. Then, a real-time implementation of above formula is achieved keypoint-wise by simply propagating it across frames via the moving average trick according to the following update scheme: $\hat{D}_{KL}(t) = \begin{cases} D_{KL}(P_1||P_0), \text{if } t = 1 \\ \alpha D_{KL}(P_t||P_{t-1}) + (1-\alpha)\hat{D}_{KL}(t-1), \forall t = 2, \ldots \end{cases}$, where $\alpha \in (0, 1)$ stands for a real parameter that we have experimentally tuned to 0.1. An example of temporal textural variation for three classes of SIFT keypoints (wave, wake, and object) in a 4K maritime video is presented in Fig. 2. One can notice, indeed, that the temporal KLpD profile of object keypoints quickly decreases, whereas, its wave and wake counterparts keep higher values over time.



(a) (b)

Fig. 3. Results of texture discrimination in a maritime video. In green: image blocks ranked as background; in blue: image blocks ranked as object. (Color figure online)

The final stage, then, consists, in merely classifying every tracked SIFT keypoint either as object or background based on its estimated value of $\hat{D}_{KL}(t)$. Such a classification algorithm, thus, attempts to divide the set of real $(1D)$ points consisting of all the values of $\hat{D}_{KL}(t)$ for each SIFT keypoint into two separate clusters (i.e.; as background *vs* object). This is efficiently achieved by means of the expectation maximization (EM) algorithm by computing the mixture of two Gaussians which best fits the set of keypoint-wise values of $\hat{D}_{KL}(t)$.

4 Spatial Texture Discrimination

Since, the above rigidity based approach only yields sparse object regions, one, moreover, needs perform some image processing in order to identity full object zones. As mentioned earlier, this is achieved in the present approach based on the analysis of texture. The intuition behind the latter consists in that objects exhibit much richer textural features (e.g.; discontinuities, etc) independently of the scale, whereas, dynamic background is generally characterized with monotonous

texture, and hence, poorer textural variation. In this paper, we have used the RFA descriptor [10] which has the advantage to be invariant against rotation and translation. The latter is computed in small overlapping image regions for capturing textural features across an image. Next, in order to capture textural variation, we perform a PCA on the matrix consisting of a column-wise alignment of the RFA vectors in a given image region in the goal of extracting the main directions of textural variation. Our experiments have demonstrated that, generally, only two or three principal components are enough for summing up most (i.e.; $\geq 90\%$) of the textural information. Thus, we have opted for the latter value (3), in such a way that, the output of this step consists of a $3D$ vector containing the three most significant eigen values of the PCA matrix. The latter is further used to feed a K-means classifier for computing two clusters, likely, corresponding to object and background, respectively.

Final maritime object extraction is merely achieved by merging image blocks– ranked as object by the texture discrimination algorithm and containing at least one stable SIFT keypoint– using the connected components algorithm. Furthermore, for the sake of accuracy and efficiency, we have implemented the latter algorithm using a multi-resolution scheme. In a nutshell, this consists in running the algorithm on a pyramid of downscaled images (by a factor of 2 with respect to each image dimension), before finally merging the results found at different scales of a 4K image for extracting full object zones. An example of the obtained results of the latter approach on a 4K maritime video is presented in Fig. 3.

5 Experimental Work

The proposed method is implemented in C++ using the OpenCV library[1], and runs in near real-time for 4K videos at an average rate of 6 images per min on an Intel CPU architecture (i5 2.2Gh).

We have chosen to show in Fig. 4 results of the proposed method using a maritime videos we have captured using a 4K fix security camera, and that we have named Video 4, throughout this experimental section. One can observe that, despite the fact that the camera is not moving, in contrast to our method, the KNN method produces awful oversegmentations resulting in many false detections. This can be explained by the fact that the mixture of Gaussians model is not well suited to maritime background.

We have also conducted a comparative study of the proposed approach against some existing well known videosurveillance approaches [25] [24], *Lin et al.* [10], and *Moo Yi et al.* [23]. However, since, we have not found any publicly available 4K video datasets with ground truth, we have only been able to test the proposed method using the following publicly available lower resolution videos:

[1] http://opencv.org/downloads.html.

– *Singapore Maritime Dataset* (*SMD*) [17]: this dataset provides several onshore RGB video sequences captured with a 70*D* Canon camera (1080×1920 px.) and showing vessels evolving in a maritime scene,
– *UCSD Background Subtraction* dataset [12]: this dataset provides several videosurveillance sequences (344×224 px.) presenting dynamic backgrounds.

We have chosen to use the F-measure of which formula is given by

$$F\text{-}measure = \frac{1}{n}\sum_{i=1}^{n} 2 * \frac{\mathrm{Prec}_i * \mathrm{Rec}_i}{\mathrm{Prec}_i + \mathrm{Rec}_i}$$

where i stands for the frame index, $Prec_i = TP_i/TP_i + FP_i$ and $Rec_i = TP_i/TP_i + FN_i$ where TP, FP and FN stand for the number of true positives, the number of true negatives and the number of false negatives, respectively.

The comparison results are summed up in Table 1 (using our 4K video datasets), Tables 2 and 3 (using publicly available video datasets).

One can observe that our method achieves best performances in terms of the F-measure which, in some sense, means that it achieves the best false positive *vs.* false negative trade-off. This can be explained by the fact that, in contrast to other existing approaches, the proposed approach in this paper is based on the useful rigid nature of maritime objects as opposed to non-rigidity of maritime background, moreover, such a rigidity hypothesis turns out to be particularly useful for eliminating wake regions.

Table 1. F-measure results using 4K videos.

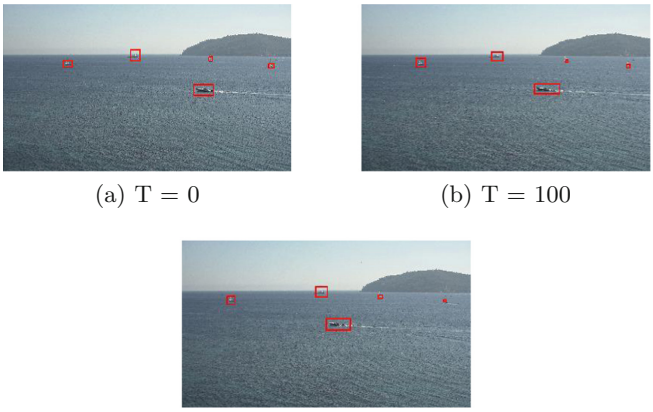
Seq	Frame number	Our method	KNN
Video1	94	**0.81**	6.27e−7
Video2	95	**0.93**	1.55e−7
Video3	64	**0.6**	1.5e−6
Video4	95	**1**	5.7e−6

Table 2. F-measure results using the Singapore maritime dataset

Seq	Num. frame	Our method
MVI_1610_VIS	537	0.727642
MVI_1646_VIS	514	0.656212

Table 3. F-measure results using the UCSD dataset.

Seq.	bottle	jump	skiing	birds
Num. frame	25	75	105	65
Our method	**0.76**	**0.64**	**0.33**	**0.36**
RFA	0.58	0.63	0.18	0.1
KNN [25]	0.54	0.39	0.08	0.048
MOG2 [24]	0.259	0.27	0.02	0.015
Dual-mode SGM	0.004	0.0032	0.027	0.09



(a) T = 0 (b) T = 100

(c) T = 200

Fig. 4. Results of the proposed approach using Video 4.

6 Conclusion

We have described a novel spatiotemporal approach to maritime target recognition approach using 4K maritime videos. The approach is mainly based on the notion of object rigidity and the property that, generally, maritime objects are texturally richer than maritime background. Moreover, for the sake of computational efficiency, first, we have proposed a parsimonious implementation of the rigidity measure by assessing the temporal deformation of the SIFT descriptor at key video locations. Second, we have developed a multi-resolution scheme for extracting full object zones based both on rigidity analysis and spatial discrimination. The present method has been implemented on a CPU architecture and achieves near real-time performances, however, the former is highly parallelizable. Thus, as future work, we will develop a GPU implementation of the present approach for achieving real-time maritime video-surveillance using 4K videos.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
2. Bechar, I., Lelore, T., Bouchara, F., Guis, V., Grimaldi, M.: Object segmentation from a dynamic background using a pixelwise rigidity criterion and application to maritime target recognition. In: ICIP, pp. 363–367 (2014)
3. Bechar, I., Lelore, T., Bouchara, F., Guis, V., Grimaldi, M.: Toward an airborne system for near real-time maritime video-surveillance based on synchronous visible light and thermal infrared video information fusion. In: OCOSS (2013)
4. Cruz, G., Bernardino, A.: Aerial detection in maritime scenarios using convolutional neural networks. In: Blanc-Talon, J., Distante, C., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2016. LNCS, vol. 10016, pp. 373–384. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48680-2_33
5. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-X_50
6. Gallego, J., Pardas, M., Haro, G.: Bayesian foreground segmentation and tracking using pixel-wise background model and region based foreground model. In: ICIP 2009, vol. 50, pp. 566–571 (2015)
7. Grimaldi, M., Bechar, I., Lelore, T., Guis, V., Bouchara, F.: An unsupervised approach to automatic object extraction from a maritime video scene. In: IPTA, pp. 378–383 (2014)
8. Liu, Y., Cui, H.Y., Kuang, Z., Li, G.Q.: Ship detection and classification on optical remote sensing images using deep learning. In: ITM Web Conference, vol. 12, p. 05012 (2017)
9. Leggat, J., Litvak, T., Parker, I., Sinha, A., Vidalis, S., Wong, A.: Study on persistent monitoring of maritime, great lakes and St. lawrence seaway border regions. Contract report DRDC CSS CR, 2011–2028 (2011)
10. Lin, S.C.F., Wong, C.Y., Jiang, G., Rahman, M.A., Kwok, N.M.: Radial fourier analysis (RFA) image descriptor. In: 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 814–819, August 2014. https://doi.org/10.1109/FSKD.2014.6980942
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
12. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 171–177 (2010)
13. Mittal, A., Paragios, N.: Motion-based background subtraction using adaptive kernel density estimation. In: CVPR, pp. 302–309 (2004)
14. Narayana, M., Hanson, A., Learned-Miller, E.: Coherent motion segmentation in moving camera videos using optical flow orientations. In: ICCV, pp. 1577–1584 (2013)
15. Oneata, D., Revaud, J., Verbeek, J., Schmid, C.: Spatio-temporal object detection proposals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 737–752. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_48
16. Pires, N., Guinet, J., Dusch, E.: ASV: an innovative automatic system for maritime surveillance. Navigation **58**(232), 1–20 (2010)
17. Prasad, D.K., Rajan, D., Rachmawati, L., Rajabally, E., Quek, C.: Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey. IEEE Trans. Intell. Transp. Syst. **18**, 1993–2016 (2017)

18. Rhodes, B.J., et al.: SeeCoast: persistent surveillance and automated scene understanding for ports and coastal areas. In: SPIE, vol. 6578, no. 1, p. 65781 (2007)
19. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the 2011 International Conference on Computer Vision, ICCV 2011, pp. 2564–2571. IEEE Computer Society (2011)
20. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV, pp. 1219–1225 (2009)
21. Smith, A., Teal, M.: Identification and tracking of marine objects in nearinfrared image sequences for collision avoidance. In: 7th International Conference on Image Processing and Its Applications, pp. 250–254 (1999)
22. Vidal, R., Soatto, S., Sastry, S.: Segmentation of dynamic scenes from the multibody fundamental matrix. In: Proceedings of the Workshop on Analysis of Dynamic Scenes (2002)
23. Moo Yi, K., Yun, K., Wan Kim, S., Jin Chang, H., Young Choi, J.: Detection of moving objects with non-stationary cameras in 5.8ms: bringing motion detection to your mobile device. In: CVPR Workshops, pp. 27–34. IEEE Computer Society (2013)
24. Zivkovic, Z.: Improved adaptive Gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004. vol. 2, pp. 28–31, August 2004
25. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. **27**, 773–780 (2006)