Fast Image-Based Obstacle Detection From Unmanned Surface Vehicles

Matej Kristan, *Member, IEEE*, Vildana Sulić Kenk, Stanislav Kovačič, *Member, IEEE*, and
Janez Perš, *Member, IEEE*

Abstract—Obstacle detection plays an important role in unmanned surface vehicles (USVs). The USVs operate in a highly diverse environments in which an obstacle may be a floating piece of wood, a scuba diver, a pier, or a part of a shoreline, which presents a significant challenge to continuous detection from images taken on board. This paper addresses the problem of online detection by constrained, unsupervised segmentation. To this end, a new graphical model is proposed that affords a fast and continuous obstacle image-map estimation from a single video stream captured on board a USV. The model accounts for the semantic structure of marine environment as observed from USV by imposing weak structural constraints. A Markov random field framework is adopted and a highly efficient algorithm for simultaneous optimization of model parameters and segmentation mask estimation is derived. Our approach does not require computationally intensive extraction of texture features and comfortably runs in real time. The algorithm is tested on a new, challenging, dataset for segmentation, and obstacle detection in marine environments, which is the largest annotated dataset of its kind. Results on this dataset show that our model outperforms the related approaches, while requiring a fraction of computational effort.

Index Terms—Autonomous surface vehicles, Gaussian mixture models, Markov random fields (MRFs), obstacle-map estimation.

I. INTRODUCTION

OBSTACLE detection is of central importance for lower-end small unmanned surface vehicles (USV) used for patrolling coastal waters (see Fig. 1). Such vehicles are typically used in perimeter surveillance, in which the USV travels along a preplanned path. To quickly and efficiently respond to the challenges from highly dynamic environment, the USV requires an onboard logic to observe the surrounding, detect potentially dangerous situations, and apply proper route modifications. An important feature of such vessel is the ability to detect an obstacle at sufficient distance and react by replanning its path to avoid collision. The primary type of obstacle in
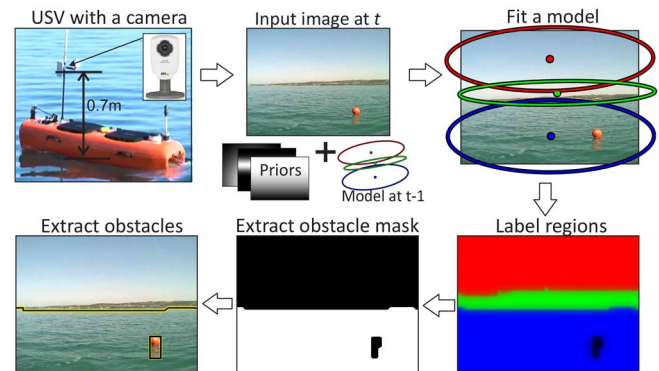
Fig. 1. Our approach to obstacle image-map estimation.

this case is the shoreline itself, which can be avoided to some extent (although not fully) by the use of detailed maps and the satellite navigation. Indeed, Heidarsson and Sukhatme [1] proposed an approach that utilizes an overhead image of the area obtained from Google maps to construct a map of static obstacles. But such an approach cannot handle a more difficult class of dynamic obstacles that do not appear in the map (e.g., boats, buoys, and swimmers).

A small USV requires ability to detect nearby and distant obstacles. The detection should not be constrained to objects that stand out from the water, but should also detect flat objects, like debris, emerging scuba divers, etc. Operation in shallow waters and marinas constrains the size of USV and prevents the use of additional stabilizers. This puts further constraints on the weight, power consumption, and types of sensors and their placement. Cameras are therefore becoming attractive sensors for use in low-end USVs due to their cost-, weight-, and power-efficiency and a large field of view coverage. This presents a challenge for the development of highly efficient computer vision algorithms tailored for obstacle detection in a challenging environments that the small USVs face. In this paper, we address this challenge by proposing a segmentation-based algorithm for obstacle-map estimation that is derived from optimizing a new, well-defined graphical model and runs at over 70 frames/s in MATLAB on a single core machine.

A. Related Work

The problem of obstacle detection has been explicitly or implicitly addressed previously in the field of unmanned

ground vehicles (UGVs). In a trail-following application, Rasmussen *et al.* [2] used an omnidirectional camera to detect trail as a region that is most contrasted to its surrounding, however, dynamic obstacles are not addressed. Montemerlo *et al.* [3] and Dahlkamp *et al.* [4] addressed the problem of low-proximity road detection with laser scanners by bootstrapping color segmentation with the laser output. The proximal road points are detected by laser, projected to camera, and used to learn a Gaussian mixture model which in turn used to segment the rest of the image captured by the camera. Combined with horizon detection of Ettinger *et al.* [5], this approach significantly increases the distance at which the obstacles on the road can be detected. Alternatively, Lu and Rasmussen [6] casted the obstacle detection as a labeling task in which they employ a bank of pretrained classifiers to 3-D point clouds and a Markov random field (MRF) to account for the spatial smoothness of the labeling.

Most UGV approaches for obstacle detection explicitly or implicitly rely on ground plane estimation from range sensors and are not directly applicable to aquatic environments encountered by USV. Rankin and Matthies [7] proposed a specific body-of-water detector in wide open areas from a UGV using a monocular color camera. Their detector assumes that, in an undisturbed water surface, a change in saturation-to-brightness ratio across a water body from the leading to trailing edge is uniform and distinct from other terrain types. They apply several *ad hoc* processing steps to gradually grow the water regions for the initial candidates and apply a sequence of preset thresholds to remove spurious false detections of water pools. However, their method is based on the undisturbed water surface assumptions, which is violated in coastal and open water applications. Scherer *et al.* [8] proposed a water detection algorithm using a stereo bumblebee camera, inertial measurement unit (IMU)/GPS, and rotating laser scanner for navigation on a river. Their system extracts color and texture features over blocks of pixels and eliminates the sky region using a pretrained classifier. A horizon line, obtained from the onboard IMU, is then projected into the image to obtain samples for learning a color distribution of the regions below and above horizon, respectively. Using these distributions, the image is segmented and results of the segmentation are used in turn, after additional postprocessing steps, to train a classifier. The trained classifier is fused with a classifier from the previous frames and applied to the blocks of pixels to detect the water region. This system relies heavily on the quality of hardware-based horizon estimation, accuracy of pretrained sky detector, and the postprocessing steps. The authors report that the vision-based segmentation is not processed on board, but requires special computing hardware, which makes it below a real-time segmentation at constrained processing power typical for small USVs.

Some of the standard range sensor modalities for autonomous navigation in maritime environments include ladar [7], radar [9], and sonar [10]. Range scanners are known to poorly discriminate between water and land in the far field [11], suffer from angular resolution and scanning rate limitations, and poorly perform when the beam's incidence angle is not oblique with respect to the water

surface [12], [13]. Several researchers have thus resorted to cameras (see [7], [13]–[17]), for obstacle and moving object detection instead. To detect dynamic objects in harbor, Socek *et al.* [14] assumed a static camera and apply background subtraction combined with motion cues. However, background subtraction cannot be applied to a highly dynamic scenes encountered on a moving USV. Huntsberger *et al.* [17] attempted to address this issue using stereo systems, but require large baseline rigs that are less appropriate for small vessels due to the increased instability and limit processing of near-field regions. Santana *et al.* [13] applied fusion of Lukas–Kanade local trackers with color oversegmentation and a sequence of *k*-means clusterings on texture features to detect water regions in videos. Alternatively, Fefilatyev and Goldgof [15] and Wang *et al.* [16] applied a low-power solution using a monocular camera for obstacle detection. They first detect the horizon line and then search for a potential obstacle in the region below the horizon. A fundamental drawback of these approaches is that they approximate the edge of water by a horizon line and cannot handle situations in coastal waters, close to the shoreline or in marina. At that point, the edge of water does not correspond to the horizon anymore and can be no longer modeled as a straight line. Such cases call for more general segmentation approaches.

Many unsupervised segmentation approaches have been proposed in literature. Khan and Shah [18] used optical flow, color, and spatial coordinates to construct features which are used in single Gaussians to segment a moving object in video. Nguyen and Wu [19] proposed Student-*t* mixture models for robustifying segmentation. Improved segmentation can be achieved by applying Bayesian regularization scheme in Gaussian mixture models, however, care has to be taken at initialization [20]. Felzenszwalb and Huttenlocher [21] have proposed a graph-theoretic clustering to perform segmentation of color images into visually-coherent regions. The assumption that the neighboring pixels likely belong to the same class is formally addressed in the context of MRFs [22], [23]. By constraining the solutions of the segmentations to mimic high-level semantics of urban scenes, Felzenszwalb and Veksler [24] proposed a three-strip segmentation algorithm that can be implemented by a dynamic program. Wojek and Schiele [25] have extended the conditional random fields with dynamic models and perform the inference for object detection and labeling jointly in videos. The random field frameworks [26] have proven quite successful for addressing the semantic labeling tasks and recently Kontschieder *et al.* [27] have showed that structural priors between classes further improve the labeling. Alternative schemes that avoid applying an MRF to enforce spatial consistency have been proposed [28], [29]. The approaches like [25] used high-dimensional features composed of color and texture at multiple scales and object-class specific detectors to segment the images and detect the objects of interest. In our scenarios, the possible types of dynamic obstacles are unknown and vary significantly in appearance. Thus, object-class specific detectors are not suitable. Several bottom-up graph-theoretic approaches have been proposed for unsupervised segmentation [30]–[33]. Recently, Alpert *et al.* [32] have proposed an approach that starts from
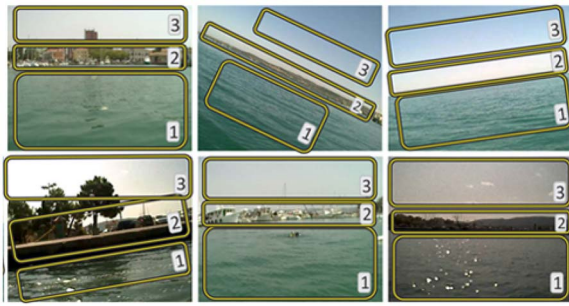
Fig. 2. Images captured from the USV split into three semantically different regions.

a pixel level and gradually constructs visually-homogenous regions by agglomerative clustering. They achieved impressive results on a segmentation dataset in which an object was occupying a significant portion of an image. Unfortunately, since their algorithm incrementally merges regions, it is too slow for online application even at moderate image sizes. An alternative to start the segmentation from pixel level is to start from an oversegmented image such that pixels are grouped into superpixels [34]. Lu *et al.* [35] applied spectral clustering to an affinity graph induced over a superpixelated image. Li *et al.* [33] have proposed a segmentation algorithm that uses multiple superpixel oversegmentations and merges their result by a bipartite graph partitioning to achieve state-of-the-art results on a standard segmentation dataset. However, no prior information is provided to favor certain types of segmentations in specific scenes.

B. Our Approach

We pursue a solution for obstacle detection that is based on concepts of image segmentation with weak semantic priors on the expected scene composition. Fig. 2 shows typical images captured from a USV. While the images significantly vary in appearance, we observe that each image can be split into three semantic regions roughly stacked one above the other, implying a structural relation between the regions. The bottom region represents the water, while the top region represents the sky. The middle component can represent either land, parked boats a haze above horizon or a mixture of these.

Our main contribution is a graphical model for structurally-constrained semantic segmentation with application to USV obstacle-map estimation. The generative model assumes a mixture model with three Gaussian components for the dominant three image regions and a uniform component for explaining the outliers, which may constitute an obstacle in the water. Weak priors are assumed on the mixture parameters and an MRF is placed over the prior as well as posterior pixel-class distributions to favor smooth segmentations. We derive an expectation maximization (EM) algorithm for the proposed model and show that the resulting optimization achieves a fast convergence at a low-computational cost, without resorting to a specialized hardware. A similar graphical model was proposed by Diplaros *et al.* [36], but their model requires a manually set variable, does not apply priors and is not derived from a single density function. Our model is applied to obstacle image-map estimation in USVs. The proposed

model acts directly on color image and does not require expensive extraction of texture-based features. Combined with efficient optimization, this results in real-time segmentation and obstacle-map estimation (several fold faster than the camera frame rate). Our approach is outlined in Fig. 1. The semantic model is fitted to the input image, after which each pixel is classified into one of the four classes. All the pixels that do not correspond to the water component are deemed to be a part of an obstacle. Fig. 1 shows a detection of a dynamic obstacle (buoy) and of a static obstacle (shoreline). Our second contribution is a marine dataset for semantic segmentation and obstacle detection, and the performance evaluation methodology. To our knowledge, this will be the largest annotated publicly available marine dataset of its kind up-to-date.

A preliminary version of our algorithm was presented in [37] and is extended in this paper on several levels. Additional discussion and related work is provided. Improved initialization of segmentation model by soft-resets of the parameters is proposed and additional details of the algorithm and the dataset are provided. In particular, the dataset capturing procedure and annotation is discussed and additional statistics of the obstacles in the dataset are provided. The experiments are extended by performance analysis with respect to the color space, the obstacle size, and the time of day driving conditions. The learning of priors used in our model is discussed in detail and the dataset is extended with training images used for estimating the priors.

Our approach is most closely related to the works in urban-scene parsing by [24] and maritime scene understanding by [8], [15], and [16]. There are notable differences between these approaches and ours. The first difference to [24] is that they only address the labeling part of the segmentation problem and require precomputed per-pixel label confidences. The second difference is that their approach produces segmentations with homogenous bottom region, which prevents detection of obstacles without further postprocessing. In contrast, our approach jointly learns the component appearance, estimates the per-pixel class probabilities, and optimizes the segmentation within a single online framework. Furthermore, learning the parameters of [24] is not as straightforward. Compared to the related water segmentation algorithms for maritime applications (see [8], [15], [16]), our approach completely avoids the need for a good horizon estimation. Nevertheless, the proposed probabilistic model is general enough to directly incorporate this information if available.

The remainder of this paper is structured as follows. Section II derives our semantic generative model. Section III presents the obstacle detection algorithm. Section IV details the implementation and learning of the priors. Section V presents the new dataset and the accompanying evaluation protocol. Section VI experimentally analyzes the algorithm and, finally, the conclusion is given in Section VII.

II. Semantic Generative Model

We consider the image as an array of measured values $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1:M}$, in which $\mathbf{y}_i \in \mathcal{R}^d$ is a d dimensional measurement, a feature vector, at the ith pixel in an image with
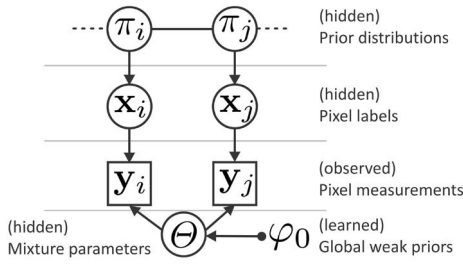
Fig. 3. Graphical model for semantic segmentation.

M pixels. As we detail in the subsequent sections, the feature vector is composed of pixel's color and image coordinates. The probability of the ith pixel feature vector is modeled as a mixture model with four components—three Gaussians and a single uniform component

$$p(\mathbf{y}_i|\Theta) = \sum_{k=1}^{3} \phi(\mathbf{y}_i|\mu_k, \Sigma_k)\pi_{ik} + \mathcal{U}(\mathbf{y}_i)\pi_{i4} \quad (1)$$

where $\Theta = \{\mu_k, \Sigma_k\}_{k=1:3}$ is the means and covariances of the Gaussian kernels $\phi(\cdot|\mu, \Sigma)$ and $\mathcal{U}(\cdot)$ is a uniform distribution. The ith pixel label x_i is an unobserved random variable governed by the class prior distribution $\pi_i = [\pi_{i1}, \ldots, \pi_{il}, \ldots, \pi_{i4}]$ with $\pi_{il} = p(x_i = l)$. The three Gaussian components represent the three dominant semantic regions in the image, while the uniform component represents the outliers, i.e., pixels that do not likely correspond to any of the three structures. To encourage segmentations into three vertically aligned semantic structures, we define a set of priors $\varphi_0 = \{\mu_{\mu_k}, \Sigma_{\mu_k}\}_{k=1:3}$ for the mean values of the Gaussians, i.e., $p(\Theta|\varphi_0) = \prod_{k=1}^{3} \phi(\mu_k|\mu_{\mu_k}, \Sigma_{\mu_k})$. To encourage smooth segmentations, the priors π_i as well as posteriors over the pixel class labels, are treated as random variables, which form an MRF. Imposing the MRF on the priors and posteriors rather than pixel labels allows effectively integrating out the labels, which leads to a well-behaved class of MRFs [36] that avoid image reconstruction during parameter learning. The resulting graphical model with priors is shown in Fig. 3.

Let $\pi = \{\pi_i\}_{i=1:M}$ denotes the set of priors for all pixels. Following [22], we approximate the joint distribution over the priors as $p(\pi) \approx \prod_i p(\pi_i|\pi_{N_i})$ and π_{N_i} is a mixture distribution over the priors of the ith pixel's neighbors, i.e., $\pi_{N_i} = \sum_{j \in N_i, j \neq i} \lambda_{ij}\pi_j$, where λ_{ij} are fixed positive weights such that for each ith pixel $\sum_j \lambda_{ij} = 1$. The potentials in the MRF are defined as

$$p(\pi_i|\pi_{N_i}) \propto \exp\left(-\frac{1}{2}E(\pi_i, \pi_{N_i})\right) \quad (2)$$

with the exponent defined as

$$E(\pi_i, \pi_{N_i}) = D(\pi_i \| \pi_{N_i}) + H(\pi_i). \quad (3)$$

The term $D(\pi_i \| \pi_{N_i})$ is the Kullback–Leibler divergence which penalizes the differences between prior distributions over the neighboring pixels (π_i and π_{N_i}), while the term $H(\pi_i)$ is the entropy defined as

$$H(\pi_i) = -\sum_{i=1}^{4} \pi_{ik} \log \pi_{ik} \quad (4)$$

which penalizes uninformative priors π_i. The joint distribution for the graphical model in Fig. 3 can be written as

$$p(\mathbf{Y}, \Theta, \pi|\varphi_0) = \prod_{i=1}^{M} p(\mathbf{y}_i|\Theta, \varphi_0)p(\Theta|\varphi_0)p(\pi_i|\pi_{N_i}). \quad (5)$$

Diplaros *et al.* [36] argue that improved segmentations can be achieved by also considering an MRF directly on the pixel posterior distributions by treating the posteriors as random variables $\mathbf{P} = \{\mathbf{p}_i\}_{i=1:M}$, where the components of $\mathbf{p}_i$ are defined as $p_{ik} = p(x_i = k|\Theta, \mathbf{y}_i, \varphi_0)$, computed by Bayes rule from $p(y_i|x_i = k, \Theta)$ and $p(x_i = k)$. We can write the posterior over $\mathbf{P}$ as $p(\mathbf{P}|\mathbf{Y}, \Theta, \pi, \varphi_0) \propto \prod_{i=1}^{M} \exp(-1/2E(\mathbf{p}_i, \mathbf{p}_{N_i}))$, where $\mathbf{p}_{N_i}$ is a mixture defined in the same spirit as π_{N_i}. The joint distribution can now be written as

$$p(\mathbf{P}, \mathbf{Y}, \Theta, \pi|\varphi_0) \propto \exp\left[\sum_{i=1}^{M} \log p(\mathbf{y}_i, \Theta|\varphi_0) - \frac{1}{2}(E(\pi_i, \pi_{N_i}) + E(\mathbf{p}_i, \mathbf{p}_{N_i}))\right]. \quad (6)$$

Due to coupling between π_i/π_{N_i} and $\mathbf{p}_i/\mathbf{p}_{N_i}$, the optimization of (6) is not straightforward. We therefore introduce auxiliary variables $\mathbf{q}_i$ and $\mathbf{s}_i$ and take the logarithm, which results in the following cost function:

$$F = \sum_{i=1}^{M}\left[\log p(\mathbf{y}_i, \Theta|\varphi_0) - \frac{1}{2}(D(\mathbf{s}_i\|\pi_i \circ \pi_{N_i}) + D(\mathbf{q}_i\|\mathbf{p}_i \circ \mathbf{p}_{N_i}))\right] \quad (7)$$

where $\circ$ is the Hadamard (component-wise) product. Note that when $\mathbf{q}_i \equiv \mathbf{p}_i$ and $\mathbf{s}_i \equiv \pi_i$, (7) reduces to (6) (ignoring the constant terms). Maximization of F can now be achieved in an EM-like fashion. In the E-step, we maximize F with respect to $\mathbf{q}_i$, $\mathbf{s}_i$, while the M-step maximizes over the parameters Θ and π. We can see from (7) that the F is maximized with respect to $\mathbf{q}_i$ and $\mathbf{s}_i$ when the divergence terms vanish, therefore, $\mathbf{s}_i^{\text{opt}} = \xi_{s_i}\pi_i \circ \pi_{N_i}$ and $\mathbf{q}_i^{\text{opt}} = \xi_{q_i}\mathbf{p}_i \circ \mathbf{p}_{N_i}$, where ξ_{s_i} and ξ_{q_i} are the normalization constants.

The M-step is not as straightforward, since direct optimization over Θ and π is intractable and we resort to maximizing its lower bound. We define $\hat{\mathbf{s}}_i = (\mathbf{s}_i + \mathbf{s}_{N_i})$ and $\hat{\mathbf{q}}_i = (\mathbf{q}_i + \mathbf{q}_{N_i})$ and by Jensen's inequality lower-bound the divergence terms as

$$-D(\mathbf{s}_i\|\pi_i \circ \pi_{N_i}) \geq \hat{\mathbf{s}}_i^T \log \pi_i$$
$$-D(\mathbf{q}_i\|\mathbf{p}_i \circ \mathbf{p}_{N_i}) \geq \hat{\mathbf{q}}_i^T \log \mathbf{p}_i \quad (8)$$

where we have ignored the terms independent of π_i and $\mathbf{p}_i$. Substituting (8) into (7) and collecting the relevant terms yields the following lower bound on the cost function (7):

$$\hat{F} = \sum_{i=1}^{M}\left[\frac{1}{2}(\mathbf{q}_i + \mathbf{q}_{N_i})^T \log(\mathbf{p}_i p(\Theta|\varphi_0)) + \frac{1}{2}(\hat{\mathbf{s}}_i + \hat{\mathbf{q}}_i)^T \log \pi_i\right]. \quad (9)$$

Differentiating (9) with respect to π_i and applying a Lagrange multiplier with the constraint $\sum_k \pi_{ik} = 1$, we see that $\hat{F}$ is maximized at $\pi_i^{\text{opt}} = (1/4)(\hat{\mathbf{s}}_i + \hat{\mathbf{q}}_i)$.

Algorithm 1 : EM for Semantic Segmentation

Require:

Pixel features $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1:M}$, priors φ_0, initial values for Θ and π.

Ensure:

The estimated parameters π^{opt}, Θ^{opt} and the smoothed posterior $\{\hat{\mathbf{q}}_{\cdot k}\}_{k=1:4}$.

Procedure:

1: Calculate the pixel posteriors $\mathbf{p}_{\cdot k}$ using the current estimates of π and Θ for all k (1).
2: Calculate the new pixel priors $\pi_{\cdot k}^{\text{opt}}$ and posteriors $\hat{\mathbf{q}}_{\cdot k}$ for all k using (12).
3: Calculate the new parameter values Θ using (10) and (11).
4: Iterate steps 1 to 3 until convergence.

Algorithm 2 : Obstacle Image Map Estimation and Obstacle Detection Algorithm

Require:

Pixel features $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1:M}$, priors φ_0, estimated model from previous time-step Θ_{t-1} and $\hat{\mathbf{q}}_{t-1}$.

Ensure:

Obstacle image map $\hat{B}_t$, water edge $\mathbf{e}_t$, detected objects $\{\mathbf{o}_i\}_{i=1:N_{\text{obj}}}$, model parameters Θ_t and $\hat{\mathbf{q}}_t$.

Procedure:

1: Initialize the parameters of Θ_t and π_t according to Section III-A.
2: Apply the Algorithm 1 to fit the model Θ_t and $\hat{\mathbf{q}}_t$ to the input data $\mathbf{Y}$.
3: Calculate the new obstacle image map $\hat{B}_t$ and for interpretation also the water edge $\mathbf{e}_t$ and the obstacles in water $\{\mathbf{o}_i\}_{i=1:N_{\text{obj}}}$.

Differentiating (9) with respect to the means and covariances of Gaussians, we obtain

$$\mu_k^{\text{opt}} = \beta_k^{-1} \left[\Lambda_k \left(\sum_{i=1}^M \hat{q}_{ik} \mathbf{y}_i^T \right) \Sigma_k^{-1} - \mu_{\mu_k}^T \Sigma_{\mu_k}^{-1} \right]^T \quad (10)$$

$$\Sigma_k^{\text{opt}} = \beta_k^{-1} \sum_{i=1}^M \hat{q}_{ik} (\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T \quad (11)$$

where we have defined $\beta_k = \sum_{i=1}^M \hat{q}_{ik}$ and $\Lambda_k = (\Sigma_k^{-1} + \Sigma_{\mu_k}^{-1})^{-1}$. An appealing property of the model (7) is that its E-step can be efficiently implemented through convolutions and Hadamard products. Recall that, the calculation of the ith pixel's neighborhood prior distribution π_{N_i} entails a weighted combination of the neighboring pixel priors π_j. Let $\pi_{\cdot k}$ be the kth component priors arranged in a matrix of image size. Then the neighborhood priors can be computed by the following convolution $\pi_{N_{\cdot k}} = \pi_{\cdot k} * \lambda$, where λ is a discrete kernel with its central element set to zero and its elements summing to one. Let $\hat{\mathbf{s}}_{\cdot k}$, $\hat{\mathbf{q}}_{\cdot k}$, and $\mathbf{p}_{\cdot k}$ be the image-sized counterparts corresponding to sets of distributions $\{\hat{\mathbf{s}}_i\}_{i=1:M}$, $\{\hat{\mathbf{q}}_i\}_{i=1:M}$, and $\{\mathbf{p}_i\}_{i=1:M}$, respectively, and let λ_1 denotes the kernel λ in which the central element is set to one. Then, the calculation of the kth component priors $\pi_{\cdot k}^{\text{opt}}$ for all pixels in the E-step can be written as

$$\hat{\mathbf{s}}_{\cdot k} = (\xi_{s \cdot} \circ \pi_{\cdot k} \circ (\pi_{\cdot k} * \lambda)) * \lambda_1$$
$$\hat{\mathbf{q}}_{\cdot k} = (\xi_{q \cdot} \circ \mathbf{p}_{\cdot k} \circ (\mathbf{p}_{\cdot k} * \lambda)) * \lambda_1$$
$$\pi_{\cdot k}^{\text{opt}} = (\hat{\mathbf{s}}_{\cdot k} + \hat{\mathbf{q}}_{\cdot k})/4. \quad (12)$$

The EM procedure for fitting our generative model to the input image is summarized in Algorithm 1.

III. OBSTACLE DETECTION

We formulate the obstacle detection as a problem of estimating an image obstacle map, i.e., determining the pixels in the image that correspond to the sea while all the remaining pixels represent the potential obstacles. We therefore first fit our semantic model from Section II to the input image and estimate the smoothed *a posteriori* probability distribution $\hat{\mathbf{q}}_{ik}$ across the four semantic components for each pixel.

An ith pixel is classified as water if the corresponding posterior $\hat{\mathbf{q}}_{ik}$ reaches maximum for the water component among all four components. In our setting, the component indexed by $k = 1$ corresponds to water region, which results in the labeled image B with the ith pixel label b_i defined as

$$b_i = \begin{cases} 1; & \arg\max_k \hat{\mathbf{q}}_{ik} = 1 \\ 0; & \text{otherwise.} \end{cases} \quad (13)$$

Retaining only the largest connected region in the image B results in the current obstacle image map $\hat{B}_t$. All blobs of nonwater pixels within the connected water region are proclaimed as potential obstacles in the water. This is followed by a nonmaxima suppression stage which merges detection that are located in close proximities (e.g., due to object fragmentation) to reduce multiple detections of the same obstacle. The water edge is extracted as the longest connected outer edge of the connected region corresponding to the water. The obstacle detection is summarized in Algorithm 2 and visualized in Fig. 1.

A. Initialization

Algorithm 1 requires initial values for the parameters Θ_t and π_t. At the first frame, when no other prior knowledge exists, we construct the initial distribution by vertically splitting the image into three regions {0, 0.2}, {0.2, 0.4}, and {0.6, 1}, written in proportions of the image height (see Fig. 4). A Gaussian is computed from each region, thus forming the observed components $\Theta_{\text{obs}} = \{\mu_{\text{obs}k}, \Sigma_{\text{obs}k}, w_{\text{obs}k}\}_{k=1:3}$. The prior over all pixels is initialized to equal probabilities for the three components, while the prior on the uniform component is set to a low constant value (see Section IV). These parameters are used to initialize the EM in Algorithm 1.

The shape for the vertical splits in Fig. 4 should ideally follow the position (and inclination) of true horizon for optimal initialization of the parameters. An estimate of the true horizon depends on the camera placement and can ideally be obtained externally from an IMU sensor, but the per-frame IMU measurements are not available in the dataset that is used in our evaluation (Section V). Therefore, an assumption
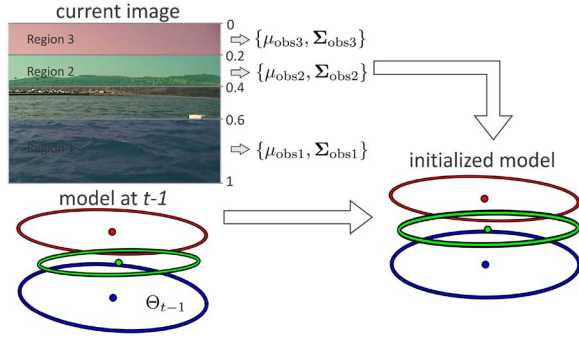
Fig. 4. Illustration of image partitioning for extraction of Θ_{obs} components and combination with the model from the previous time-step Θ_{t-1} for initialization of the EM.
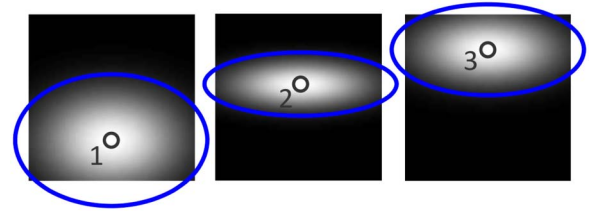

Fig. 5. Visualization of the spatial part of the Gaussians in the weak priors over our three semantic components. From left to right: bottom-most, middle, and top-most component prior.

is made that the horizon, as well as edge of water, is usually located within the region $\{0.4, 0.6\}$ of image height, which is the reason this region is excluded from computation of parameter initial values. Making no further assumptions regarding the proportion between components in the final segmentation, equal regions (2 and 3) are used to initialize the parameters of components 2 and 3. The assumption on region splitting is often violated in our dataset from Section V due to boat inclination at turning maneuvers, due to boat tilting forward and backward, and since the camera might not have been mounted to exactly the same spot in assembling the boat after transportation to the test site during several months that the dataset was taken. Nevertheless, the segmentation algorithm is robust enough to handle the nonideal initializations as long as there are no extreme deviations, like the boat toppling or riding on extremely high waves (the small coastal USVs are actually not even designed to physically endure these extreme weather conditions).

During the USVs operation, we can exploit the continuity of sequential images in the videostream by using the parameter values of the converged model from the previous time-step for initialization of the EM algorithm in the current time-step. To reduce possible propagation of errors stemming from false segmentations in the previous time-steps, a zero-order soft reset is applied in the initialization of the EM in each time-step. In particular, the EM is initialized by merging the Θ_{obs} with Θ_{t-1}. The parameters of the kth component, $\{\mu_{\text{init}k}, \Sigma_{\text{init}k}\}$, are initialized by forming a weighted two-component mixture model from the kth components in Θ_{obs} and Θ_{t-1}, and approximating them by a single component by matching to first two moments of the distributions [38]–[40]. The weights α and $1 - \alpha$ for Θ_{t-1} and Θ_{obs}, respectively, can be used to balance the contribution of each component. The priors π_t over the pixels are initialized by the smoothed posterior $\hat{\mathbf{q}}_{\cdot k}$ from the previous time-step. The initialization is sketched in Fig. 4.

IV. IMPLEMENTATION DETAILS

In our application, the measurement at each pixel is encoded by a five-dimensional feature vector $\mathbf{y}_i = [i_x, i_y, i_{c1}, i_{c2}, i_{c3}]$, where (i_x, i_y) are the ith pixel coordinates and the (i_{c1}, i_{c2}, i_{c3}) are the pixel's color channels. We have determined in a preliminary study that we achieve sufficiently good obstacle detection

by first performing detection on a reduced-size image of 50×50 pixels and then rescaling the results to the original image size. The rescaling was set to match the lower scale of objects of interest, as smaller objects do not present danger to the USV. Such approach drastically speeds up the algorithm to approximately 10 ms per frame in our experiments. The uniform distribution component in (1) is defined over the image pixels domain and returns equal probability for each pixel. Assuming that all color channels are constrained to the interval $[0, 1]$, the value of the uniform distribution is $\mathcal{U}(\mathbf{y}_i) = 1/(50^2)$ at each pixel for our rescaled image. The EM optimization requires specification of the convolution kernel λ. Note that, the only constraint on the convolution kernel is that its central element is set to zero and all elements sum to one. We use a Gaussian kernel with its central element set to zero and set the size of the kernel to 2% of image size, which results in a 3×3 pixels kernel. Recall from Section III-A that the parameter α influences the soft-reset of the parameters used to initialize the EM. In our implementation, a slightly larger weight is given to the parameters estimated at the previous time-step by setting $\alpha = 0.6$.

A. Learning the Weak Priors

The spatial components in the feature vector play a dual role. First, they enforce to some extent the spatial smoothness of the segmentation on their own. Second, they lend means to weakly constraining the Gaussian components such that they reflect the three dominant semantic image parts. This is achieved by the weak priors $p(\Theta|\varphi_0) = \prod_{k=1}^{3} \phi(\mu_k|\mu_{\mu_k}, \Sigma_{\mu_k})$ on the Gaussian means. Since the locations and shape of semantic components vary significantly with the views, we indeed select weak priors, which are estimated using the training set from our database (see Section V). Given a set of training images, the prior of the kth component is estimated by extracting the features, i.e., sets of $\mathbf{y}_i$, corresponding to the kth component from all images and fit a single Gaussian to them. Note that, in general, there is a chance that the training images might bias the horizontal location of the estimated Gaussian to the left or right part of the image. In this case, we could constrain the horizontal position of Gaussians to be in the middle of the image, however, we have observed that the components of the prior estimated from our dataset are sufficiently centered and we do not apply any such constraints.

Examples of the spatial parts of the priors estimated from the training set of the dataset presented in Section V are shown in Fig. 5. Our algorithm, as well as the learning routine,

was implemented in MATLAB—a reference code is publicly available.[1]

V. Marine Obstacle Detection Dataset

With lack of sufficiently large publicly available annotated dataset to test our method, we have constructed our own dataset, which we call the marine obstacle detection dataset (MODD). The MODD consists of 12 video sequences, providing in total 4454 fully annotated frames with resolution of 640×480 pixels. The dataset is made publicly available along with the annotations and MATLAB evaluation routines from the MODD homepage.[2]

The video sequences have been recorded from multiple platforms, most of them from the small 2.2 m USV[3] (see Fig. 1). The USV was developed by Harpha Sea, d.o.o. Koper, and is based on catamaran hull design and powered by electrical, LiPo battery-powered, steerable thrust propeller. It can reach the maximum speed of 2.5 m/s and has extremely small turn radius. Steering and route planning are handled by ARM-powered microcontroller (MCU) with redundant power supply. For navigation, the MCU relies on microelectromechanical inertial navigation unit (MEMS IMU), solid-state digital compass, and differential GPS. USV has two different communication channels to the shore (high- and low-bandwith) and its mission can be programmed remotely. An Axis 207 W camera was placed on the USV approximately 0.7 m above the water surface, looking in front of the vessel, with an approximately 55° field of view. Camera has been set up to automatically adjust to the variations in lighting conditions. Since the boat was being reassembled between the runs over several months, the placement of the camera varies slightly across the dataset.

The video sequences have been acquired in the Gulf of Trieste, specifically in the port of Koper, Slovenia (Fig. 8), over a period of months at different times of day under different weather conditions. The USV was manually operated by a human pilot and effort was made to simulate realistic navigation, including threats of collision. The pilot was instructed to deliberately drive in a way to simulate situations in which an obstacle might present a danger to the USV. This includes obstacle being very close to the boat as well as situations in which the boat was heading straight toward an obstacle for a number of frames.

The first ten videos in the dataset are meant for evaluation of the obstacle-map estimation algorithms under normal conditions. These videos still vary quite significantly between each other and simulate conditions under which the USV is expected to operate. We thus term these ten videos as normal conditions and we show some examples of these videos in the first ten images from Fig. 6. The last two videos were meant for analysis in situations in which the boat is directly facing the sun. This causes extreme changes in the automatic shutter and camera setting, resulting in significant changes of contrast and color of all three semantic components. Facing the sun also generates significant amount of fragmented glitter,
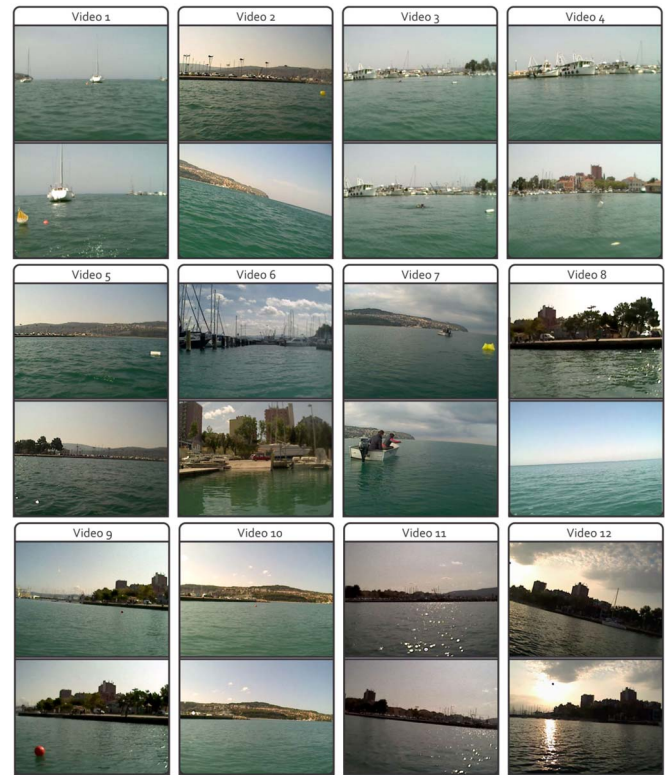

Fig. 6.   Examples of images taken from the videos in the MODD. The first ten videos are for normal conditions, while the last two depict extreme conditions. For each video, we show two images for better impression of the video content.

while sometimes it shows up as a larger, fully connected region of the reflected sun. We thus denote these last two videos as extreme conditions. Some examples are shown in the last two images of Fig. 6.

Each frame is annotated manually by a polygon denoting the edge of water and bounding boxes are placed on large obstacles (those that straddle the water edge) and small obstacles (those that are fully surrounded by water). See Fig. 8 for illustration. The annotation was made by a human annotator and all annotations on all images of the dataset were later verified by an expert. To allow a fast overview of the annotations by the potential users of the dataset, the dataset provides a rendered video with annotations overlay, for each test sequence in the dataset—these videos are included as part of the dataset and available from the dataset homepage as well.

In the next section, some general statistics of the dataset are provided. The boat was driving within 200 m from the shore, and most of the depicted obstacles are in this zone. Out of 12 image sequences in the dataset, nine contain either large or small obstacles, one contains only annotated sea edge, and two contain glitter annotations, sea edge annotations, and no objects. The number of objects per frame is exponentially distributed with the average 1.1 and variance 1.23. The distribution of the annotated size of small and large obstacles is shown in Fig. 7. For the algorithms that require training or validation of their parameters, we have compiled a collection of 20 images in which we manually annotated the pixels corresponding to three semantic components. Fig. 10 shows some examples of images and the corresponding annotations.

---

[1]http://www.vicos.si/Research/UnmannedSurfaceVehicles

[2]http://www.vicos.si/Downloads/MODD

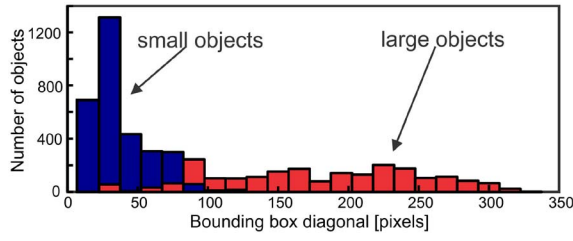[3]A video of our USV is available online from the MODD homepage.

Fig. 7. Distributions of the floating obstacle sizes labeled as large and small shown in red and blue, respectively.
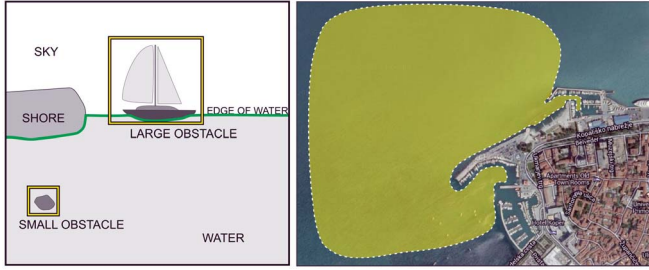

Fig. 8. Left: scene representation used for evaluation. Right: dashed outline denotes the region in the coastal waters of Koper, Slovenia (Gulf of Trieste), where the dataset was captured.
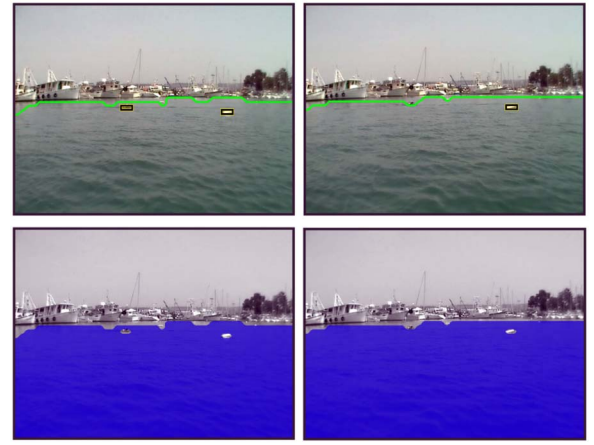

Fig. 9. Images with a scuba diver emerging from the water just at the observed water edge. Top: water edge and detected obstacles in the water. Bottom: water mask. Left: estimated water edge runs above the diver and the scuba diver is explicitly detected. Right: edge runs below the scuba diver, which prevents explicit detection, even though the region corresponding to the scuba diver is in fact detected as a part of the obstacle.
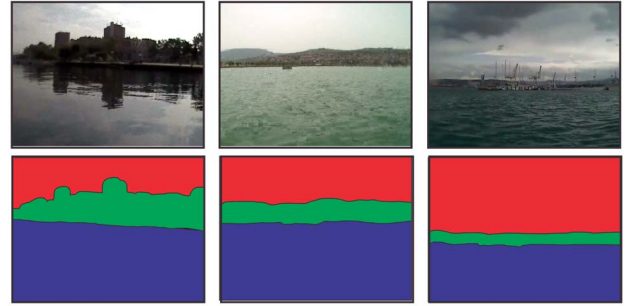

Fig. 10. Examples of training images along with their manual label masks. The blue, green, and red color corresponds to the labels for bottom, middle, and the top semantic component, respectively.

A. Evaluation Protocol

The evaluation protocol is designed to reflect the two distinct challenges that the USVs face: 1) the water edge (shoreline/horizon) detection and 2) obstacle detection. The former is measured as the root mean square error (RMSE) of the water edge position (*Edg*), and the latter is measured via the efficiency of small object detection, expressed as precision (*Prec*), recall (*Rec*), *F*-score (*F*), and the average number of false positives (FPs) per frame (*aFP*).

To evaluate RMSE in water edge position, ground truth annotations were used in the following way. A polygon, denoting the water surface was generated from water edge annotations. Areas, where large obstacles intersect the polygon, were removed. Note that, given the scene representation, shown in Fig. 8, one cannot distinguish between large obstacles (e.g., large ships) and stationary elements of the shore (e.g., small piers). This way, a refined water edge was generated. For each pixel column in the full-sized image, a distance between water edge, as given by the ground truth and as determined by the algorithm, was calculated. These values are summarized into a single *Edg* value by averaging across all frames and videos.

The evaluation of object detection follows the recommendations from the PASCAL VOC challenges by Everingham *et al.* [41], with small, application-specific modification: all small obstacles (provided as a ground truth or detected) that are closer to the annotated water line than 5% of the image height, are discarded prior to evaluation on each frame. This aims to address situations where a detection may oscillate between fully water-enclosed obstacle, and the "dent" in the shoreline, resulting in false negatives (FNs). Fig. 9 shows an example with two images of a scuba diver emerging from the water. Note that in both images, the segmentation successfully labeled the scuba diver as an obstacle. But in the

left-hand image we obtain an explicit detection, since the estimated water edge runs above the scuba diver. In the right-hand image, the edge runs below the scuba diver and we do not get explicit detection, even though the algorithm successfully labeled the scuba diver's region as being an obstacle. Note that, the proposed treatment of near-edge detections/ground-truths is also consistent with the problem of obstacle avoidance— the USV is concerned primarily with the avoidance of the obstacles in its immediate vicinity. In counting FPs, true positives (TP), and FNs, we follow the methodology of PASCAL VOC, with the minimum overlap set to 0.3. FP, TP, and FN were used to calculate *Prec*, *Rec*, *F*, and *aFP*.

VI. EXPERIMENTS

In this section, we will denote our obstacle image-map estimation method (Algorithm 2) as the semantic-segmentation model (SSM). The experimental analysis was split into three parts. In the first part, we evaluate the influence of the different color spaces on the SSMs performance. In the second and third part, we analyze how various elements of SSM affect its performance and compare it to the alternative methods. All experiments were performed on a desktop PC with 3.06 GHz Intel Xeon E5-1620 CPU in a single thread in MATLAB.

colorspace	Edg[pix]	Prec	Rec	F	aFP
RGB	10.7(5.8)	0.874	0.756	0.806	0.039
HSV	12.7(8.3)	0.821	0.688	0.738	0.041
Lab	**9.3**(5.1)	0.878	0.768	0.815	0.039
YCrCb	9.5(5.5)	**0.885**	**0.772**	**0.819**	**0.039**

A. Influence of the Color Space

The aim of the first experiment was to evaluate how the different colorspaces affect the segmentation performance. We have therefore performed experiments in which the feature vector $\mathbf{y}_i$ (Section IV) was calculated from RGB, hue, saturation, and value (HSV), Lab, and YCrCb colorspace. For each of the selected colorspaces, the weak priors were learned from the training images on the MODD dataset (Section V). All experiments were performed on all 12 testing videos from the MODD dataset. The results are shown in Table I.

The results show that best performance is achieved with the YCrCb and Lab colorspace, which is not surprising, since these colorspaces are known to better cluster visually-similar colors. Similar is true for the HSV space, but that space suffers from the circular property of the Hue component (i.e., red color is on the left-most and right-most part of the Hue spectrum). With respect to the edge of the water estimation, the lowest error is achieved when using the Lab colorspace, while only a slightly worse performance is obtained with the YCrCb colorspace. On all other measures, the YCrCb colorspace yields best results, although comparable to the Lab colorspace. While the results are worse when using the RGB or the HSV colorspace, we note that these results do not exhibit drastically poorer performance, which speaks of a level of robustness of the SSM to the choice of the colorspace. Nevertheless, given the results in Table I, we select the YCrCb and use this colorspace in the subsequent experiments.

B. Comparison to Alternative Approaches

Given a fixed colorspace, we are left with evaluation of how much each part of our model contributes to the final performance. We have therefore also implemented two variants of our approach, which we denote by UGM and UGM$_{col}$. In contrast to SSM, the UGM and UGM$_{col}$ do not use the MRF constraints and are in this respect only mixtures of three Gaussians with priors on their means and with a uniform component. A further difference between UGM and UGM$_{col}$ was that UGM$_{col}$ ignored the spatial information in visual features and relied only on color.

Note that the SSM is conceptually similar to the Grab-cut algorithm from Rother *et al.* [42] for binary segmentation, but with distinct differences. In the Grab-cut, the user provides a bounding box roughly containing the object, thus initializing the segmentation mask. Two visual models using a GMM are constructed from this segmentation mask. One for the object and one for the background. An MRF is then constructed over the pixel grid and graph cut from Boykov *et al.* [43] is used to infer an improved segmentation mask. This procedure is then iterated until convergence. There are significant differences between the proposed SSM and the Grab-cut from [42]. In contrast to the user-provided bounding box in [42], the SSMs weak supervision comes from the initialization of the parameters from the previous time-step and from the weak priors. The second distinction is that our approach does require explicit estimation of the segmentation mask to refine the mixture model. This allows for a better propagation of uncertainty during the iteration of the algorithm, leading to improved segmentation.

To further evaluate contributions of the particular MRF optimization of our SSM, we have implemented a variant of the Grab-cut algorithm, which uses our semantic mixture model, but applies graph-cuts for optimization over the MRF. The resulting obstacle-map estimation tightly follows Algorithms 1 and 2 with a slight modification of Algorithm 1: after each epoch of the EM, we apply the graph-cut from Bagon [44] to segment the image into a water/nonwater mask. This mask is then used as in the original Grab-cut to refine the mixture model. We use exactly the same weakly-constrained mixture model as in SSM, and the YCrCb colorspace for fair comparison, and call this approach the Grab-cut model (GCM).

We have compared our approach also to the general segmentation approaches, namely the superpixel-based approach from Li *et al.* [33], SPX and a graph-based segmentation algorithm from Felzenszwalb and Huttenlocher [21], FZH.

For fair comparison, all the algorithms were executed on the 50×50 images. We have experimented with the parameters of GCM and FZH and have set them to optimal performance for our dataset. Since FZH was designed to run on larger images, we have also performed the experiments for FZH on full-sized images—we denote this variant by FZH$_{full}$. We have performed the comparative analysis separately for the normal and extreme conditions.

1) Performance Under Normal Conditions: The results of the experiments on the normal conditions part of the MODD are summarized in Table II, while Fig. 11 shows an example of typical segmentation masks from the compared algorithms. The segmentation results in these images are color coded as follows. The original image is represented only by the blue channel, manual water annotations are shown in the green channel, and algorithm-generated water segmentation is shown in the red channel. Therefore, the cyan region shows the area, which has been annotated as water, but has not been segmented as such by the algorithm (bad). The magenta region shows the area, which has not been annotated as water, but has been segmented as such by the algorithm (bad). The yellow area shows the area which has been annotated as water and has been segmented as such by the algorithm (good), and blue region shows the area which has not been annotated as water and has not been segmented as such (good). Finally, the darker band under the annotated edge of the water in all colors

TABLE II
COMPARISON OF VARIOUS METHODS UNDER NORMAL CONDITIONS.
THE RESULTS ARE GIVEN BY REPORTING AVERAGE PERFORMANCE
WITH A STANDARD DEVIATION IN BRACKETS: EDGE OF WATER
ESTIMATION ERROR IN PIXELS (*Edg*), *PrecRec*,
F MEASURE (*F*), *aFPs*, AND TIME IN [MS] (*t*)

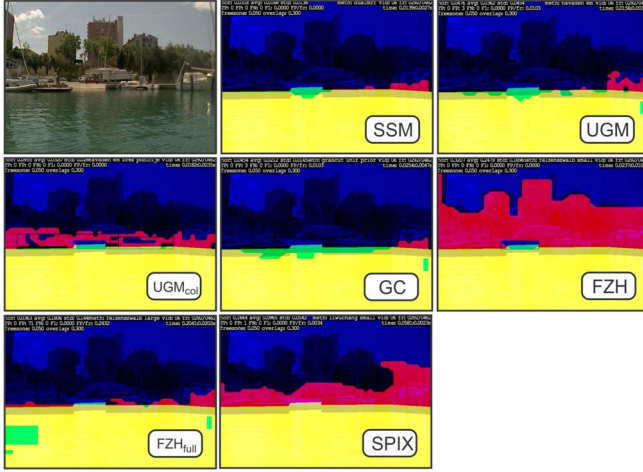|	Edg	Prec	Rec	F	aFP	t
SSM	**9.2**(4.9)	**0.885**	**0.772**	**0.819**	**0.039**	**10**(0)
GCM	10.9(5.6)	0.718	0.686	0.695	0.121	17(3)
UGM	10.5(6.1)	0.742	0.706	0.717	0.109	11(2)
UGM$_{col}$	16.4(9.0)	0.614	0.504	0.549	0.122	11(3)
FZH	90.0(65.7)	0.727	0.523	0.551	0.053	16(1)
FZH$_{full}$	34.2(41.4)	0.410	0.747	0.488	0.697	199(3)
SPX	66.4(34.7)	0.007	0.001	0.001	0.090	54(1)

Fig. 11.   Qualitative comparison of the different obstacle-map estimation approaches. The upper left-most image is the original input image, followed by the results for SSM, UGM, UGM$_{col}$, GCM, FZH, FZH$_{full}$, and SPX. This image is best viewed in color. Please see text for the description of the color codes.

shows the ignore region, where evaluation of small obstacle detection does not take place.

Recall that in contrast to the SSM, the UGM and UGM$_{col}$ do not impose an MRF constraint. In Fig. 11, this clearly leads to a poorer segmentation, resulting in FP detections of obstacles as well as significant over-segmentations. Quantitatively, the poorer performance is reflected in Table II as a lower *F*-measure, higher average number of FPs, and larger edge of the water estimation error. Compared to SSM, we observe a significant drop in detection quality of the UGM, especially precision. This speaks of importance of the local labeling constraints imposed by the MRF in the SSM. The performance further drops with UGM$_{col}$, which implies that spatial components in the feature vectors bear important information for proper segmentation as well. On the other hand, the GCM does impose an MRF, however, the segmentation is still poorer than with the SSM. We believe that the main reason for this is that the GCM applies graph-cuts to perform hard segmentation during EM updates. On the other hand, the SSM optimizes the cost function within a single EM framework, thus avoiding the need for hard segmentations during the EM steps, which leads to a better final result. By far the worst segmentation results are obtained by the FZH, FZH$_{full}$, and SPX segmentation methods.

Note that while these segmentation methods do assume some local consistency of segmentation, they still perform poorer than the SSM. The improved performance of SSM can be attributed exclusively to our formulation of the segmentation model within the graphical model from Fig. 3.

Fig. 12 shows further examples of SSM segmentation maps (the first 14 images), the spatial part of the Gaussian mixture and the detected objects in water. The appearance and texture of the water varies significantly between the various scenes, and the same is true for the other two semantic components. The images also vary in the scene composition in that the vertical position as well as the attitude of the water edge (see second row in Fig. 12) vary significantly. Nevertheless, note that the model is able to adapt well to these compositions and successfully decomposes the scene into water regions, in-water obstacles, and fairly well delineates the water edge.

Our algorithm performed (segmentation + detection) at a rate higher than 70 frames/s. Most of the processing was spent on fitting our semantic model and obstacle-map estimation (10 ms), while 4 ms was spent on the obstacle detection. For fair comparison of segmentation algorithms, we report in Table II only the times required for the obstacle-map estimation. Although note that the obstacle detection part did require more processing time for the methods that delivered poor segmentation masks with more FPs. On average, our EM algorithm in SSM converged in approximately three iterations. Note that the graph cut routine in GCM, part of SPX and the FZH were implemented in C and interfaced to MATLAB, while all the other variants were entirely implemented in MATLAB. Therefore, the computational time results for segmentations are not directly comparable among the methods, but still offer a level of insight. In terms of processing time, the SSMs segmentation was the fastest, running at 100 frames/s. The UGM$_{col}$ and UGM performed approximately as fast as SSM, followed by GCM, FZH, SPX, and FZH$_{full}$. We conclude that the SSM came out on top as the fastest method that also achieved the best detection performance as well as accuracy.

2) Performance Under Extreme Conditions: We were interested in measuring two properties of the algorithms under conditions when the boat is facing the sun. In particular, we were interested in measuring how the sun affects the edge-of-water estimation and how the glitter affects the detection. We have therefore repeated two variants of the experiments on the videos in MODD denoted by the extreme conditions (videos 11 and 12 in Fig. 6). In the first variant, we ignored any detections in the regions that were denoted as glitter regions in ground truth. In the second variant, all detections were accounted for. Note that the videos denoted as extreme conditions do not contain any objects, therefore there were no true positives and any detected object was an FP. Because of this, we present in the results (Table III) only the edge of water estimation error and the average number of FPs (by definition, both, accuracy and precision, would be zero in such case).

In terms of edge of water estimation, the UGM$_{col}$ slightly outperforms the SSM. The UGM$_{col}$ ignores the spatial information and generally oversegments the regions close to the shoreline (as seen in Fig. 11), which in this case actually reduces the error compared to SSM. The reason is that the
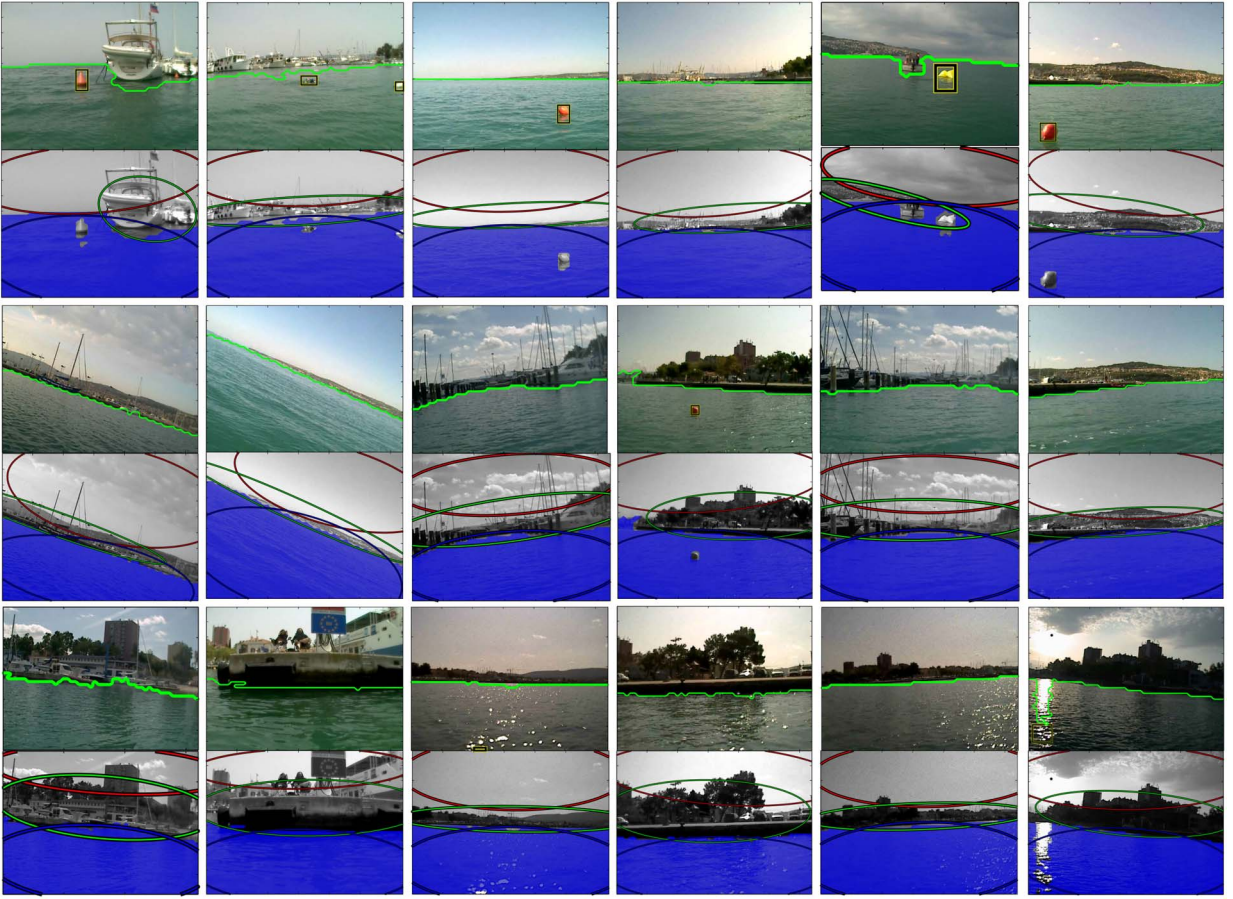
Fig. 12. Qualitative examples of water segmentation and obstacle detection. We show for each image the detected edge of the sea in green and the detected obstacle by a yellow rectangle. Below each image, we also show the spatial part of the three semantic components as three Gaussian ellipses and the portion of the image segmented as water in blue. This image is best viewed in color.

TABLE III
COMPARISON OF VARIOUS METHODS UNDER EXTREME CONDITIONS.
WE SHOW THE RESULTS FOR EDGE OF WATER ESTIMATION
ERROR *Edg* AND *aFPs* WHEN IGNORING GLITTER
AND WHEN COUNTING THE GLITTER AS FPS,
aFP_{ignore} AND aFP_{account}, RESPECTIVELY

	Edg[pix]	aFP_{ignore}	aFP_{account}
SSM	11.3(10.0)	**0.000**	0.134
GCM	16.1(15.3)	0.010	0.919
UGM	12.4(11.3)	0.007	0.932
UGM$_{\text{col}}$	**8.1**(5.1)	0.019	0.308
FZH	65.3(48.7)	0.003	0.233
FZH$_{\text{full}}$	46.4(51.6)	0.159	2.056
SPX	49.2(48.6)	0.015	**0.019**

SSM attributes the upper part of the sun reflection at the shoreline in video 11 (Fig. 6) to an obstacle instead of the water. When ignoring the glitter region, the SSM outperforms the competing methods by not detecting any FPs (zero aFP_{ignore}), while the competing methods exhibit larger values of the FPs. When also considering the glitter region, the number of FPs only slightly increases for the SSM, while this increase is considerable for the other methods. Note that in this case the SSM again significantly outperforms the other methods, except for SPX. The reason is that the SPX actually fails by

grossly oversegmenting the water region, thus assigning almost all glitter to that region. However, looking at the results of the edge estimation, we can also see that this oversegmentation actually consumes a part of the shoreline, thus leading to poor overall segmentation. Among the remaining methods, the SSM again achieves the lowest average FP rate. Given these results, we conclude that the SSM is much more robust to extreme conditions than the competing methods, while still offering good segmentation results. Some examples of segmentation with SSM are shown in the last four images of Fig. 12. Even in these harsh conditions, the model is able to interpret the scene well enough with few false obstacle detections. For more illustrative examples of our method and segmentations, please consult the additional online material at http://box.vicos.si/matejk/smc/index.htm.

3) Failure Cases: An example of conditions is which the segmentation is expected to fail is shown in the bottom-most right image of Fig. 12. In this image, the boat is facing a low-laying sun directly, which results in a large saturated glitter on the water surface. Since the glitter occupies a large region, and is significantly different from the water, it is detected as an obstacle. Such cases could be handled by image postprocessing, but at a risk of missing true detections. Nevertheless, additional sensors like compass, IMU, and sun-position model can be used to identify a detected region as a potential glitter.
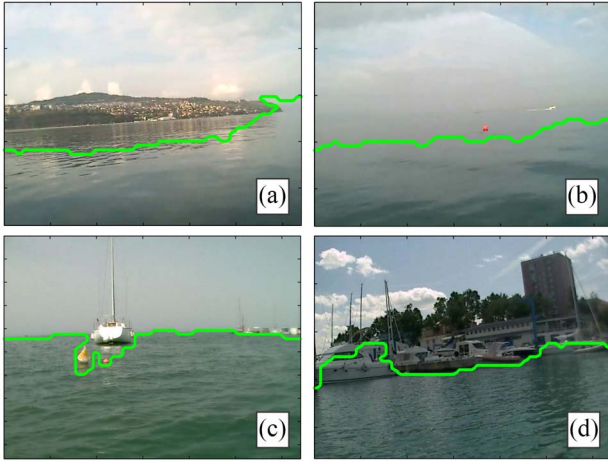
Fig. 13. Qualitative examples of poor segmentation due to a (a) strong reflection, (b) poorly visible transition from water to sky, (c) cluster of objects, and (d) obstacles visually similar to water.
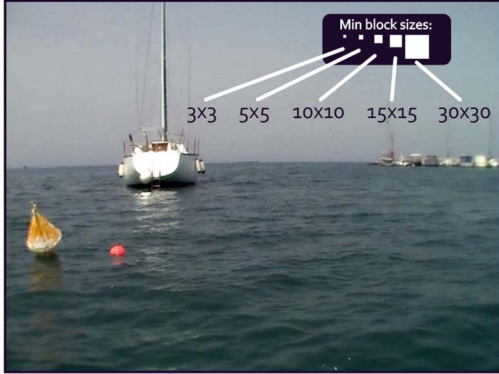

Fig. 14. Visualization of minimum thresholds used in Table IV.

To offer further insights of the constraints of the proposed segmentation, we show additional failure cases in Fig. 13. Fig. 13(a) shows failure due to a strong reflection of the land-mass in the sea, while Fig. 13(b) shows an example of failure due to blurred transition from the sea to sky. Note that in both cases, the edge of the sea is conservatively estimated, meaning that true obstacles were not mislabeled as water, but rather portions of water were labeled as obstacle. Fig. 13(c) shows an example in which several obstacles are close-by and are not detected as separate obstacles, but rather as part of the edge of water. An example of potentially dangerous mislabeling is shown in Fig. 13(d), where a part of the boat on the left is deemed visually-similar to water and is labeled as such. Note, however, that this mislabeling is corrected in the subsequent images in that video.

C. Effects of the Target Size

Note that, all obstacles may not pose equal threat to the vessel. In fact, smaller objects are likely not people and may also likely pose little threat, since they can be run over without damaging the vessel. To inspect our results in such a context, we have compiled the SSM results over all videos with respect to the minimum object's size. Any object, whether in the ground truth or in detection, was ignored if its size was

TABLE IV
RESULTS OF THE SSM$_{a\times a}$ AND RELATED APPROACHES ON ALL VIDEO SEQUENCES WITH RESPECT TO THE MINIMUM OBJECT SIZE $a \times a$. FOR REFERENCE, THE RESULTS FOR TOP-PERFORMING BASELINES ARE PROVIDED FOR 5×5, 15×15, AND 30×30 PIXELS

	Prec	Rec	F	aFP
SSM$_{0\times0}$	0.885	0.772	0.819	0.055
SSM$_{3\times3}$	0.898	0.772	0.825	0.049
SSM$_{5\times5}$	0.898	0.772	0.825	0.049
SSM$_{10\times10}$	0.898	0.773	0.826	0.049
SSM$_{15\times15}$	0.896	0.792	0.837	0.049
SSM$_{30\times30}$	0.924	0.801	0.846	0.010
GCM$_{5\times5}$	0.733	0.686	0.703	0.246
GCM$_{15\times15}$	0.731	0.701	0.701	0.246
GCM$_{30\times30}$	0.812	0.759	0.772	0.068
FZH$_{5\times5}$	0.731	0.523	0.553	0.082
FZH$_{15\times15}$	0.731	0.536	0.565	0.082
FZH$_{30\times30}$	0.779	0.577	0.614	0.038
SPX$_{5\times5}$	0.007	0.001	0.001	0.078
SPX$_{15\times15}$	0.007	0.001	0.001	0.078
SPX$_{30\times30}$	0.003	0.001	0.001	0.074

smaller than a predefined value. We also ignored any detected object that overlapped with the removed ground truth detection by 0.3. This last condition addresses the fact that some objects in the ground truth are slightly smaller than their detected size, which would generate an incorrect FP if the ground truth object was removed. Fig. 14 visualizes the applied thresholds, while the results are given in Table IV.

The results show that the detection remains high over a range of small thresholds, which speaks of a level of robustness of our approach. By increasing thresholds above 10×10 the precision as well as the recall increase the probability of detecting an FP in a given frame is drastically reduced. This means that, as the objects approach the USV and get bigger, they are increasingly reliably detected. This is also true for the sufficiently big objects that are far away from the USV. The following rule-of-thumb calculation for the big or approaching objects can be performed. Let us assume that a successful detection means any detection of a true obstacle if we detect it at least once in $N_{\text{buf}} = 3$ consecutive frames. The probability of a successful detection is

$$p_{\text{success}} = 1 - (1 - Rec)^{N_{\text{buf}}}. \tag{14}$$

If we do not apply any thresholding, we can detect any object, regardless of its size with probability 0.988. The probability of an FP occurring in any image is 0.055. By applying a small 3×3 threshold, the detection remains unchanged, but the probability of an FP occurring in a particular frame goes down to 0.05. If we chose to focus only on the objects that are at least 30×30 pixels large, then the probability of detection goes up to 0.992, and the probability of detecting an FP in any frame goes down to 0.01. It should be noted that the model in (14) assumes independence of detections over the sequence of images. While such assumptions may indeed be restrictive for temporal sequences, we still believe that the model gives a good rule-of-thumb on expected real-life obstacle detection performance of the segmentation algorithm.

VII. Conclusion

A graphical model for semantic segmentation of marine scenes was presented and applied to USV obstacle-map estimation. The model exploits the fact that scenes a USV encounters may be decomposed into three dominant visually- and semantically-distinctive components, one of which is the water. The appearance is modeled by a mixture of Gaussians and accounts for the outliers by a uniform component. The geometric structure is enforced by placing weak priors over the component means. An MRF model is applied on prior and posterior pixel-label distribution to account for the interactions across neighboring pixels. An EM algorithm is derived for fitting the model to image, which affords fast convergence and efficient implementation. The proposed model directly applies straight-forward features, i.e., color channels and pixel positions and avoids potentially slow extraction of more complex features. Nevertheless, the model is general enough to be directly applied without modifications to any other features. A straightforward approach for estimation of the weak prior model was proposed, that allows learning from a small number of training images and does not require accurate annotations. Results show excellent performance compared to related segmentation approaches and exhibits improved performance in terms of segmentation accuracy as well as speed.

To evaluate the performance and analyze our algorithm, we have compiled and annotated a new, real life coastal line segmentation dataset captured from an onboard marine vehicle camera. This is the largest dataset of its kind to date and is as such another contribution to the field of robotic vision. We have studied the effects of the colorspace selection on the algorithm's performance. We conclude that the algorithm is fairly robust to this choice, but obtains the best results at YCrCb and Lab colorspaces. The experimental results also show that the proposed algorithm significantly outperforms the related solutions. While the algorithm provides high detection rates at low FPs it does so with a minimal processing time. The speed comes from the fact that the algorithm can be implemented through convolutions and from the fact that it preforms robustly on small images. The results have also shown that the proposed method outperforms the related methods by a large margin in terms of robustness in the extreme conditions, when the vehicle is facing the sun, as well. To make this paper a reproducible research and to facilitate other researchers in comparing their work to ours, the MODD dataset is made publicly available, along with all the MATLAB evaluation routines, a reference MATLAB implementation of the presented approach and the routines for learning the weak priors.

Note that the fast performance is of crucial importance for real-life implementations on USVs, as it allows the use in onboard embedded controllers and low-cost embedded, low-resolution cameras. Our future work will focus on two extensions of our algorithm. We will explore possibilities of porting our algorithm to such an embedded sensor. Since many modern embedded devices contain GPUs, we will also explore parallelization of our algorithm by exploiting the fact that it is based on convolution operations, which can be efficiently parallelized. Our model is fully probabilistic and as such affords a principled way for information fusion (see [45]), to improve performance. We will explore combinations with additional external sensors such as inertial sensors, cameras of other modalities, and stereo systems. In particular, IMU can be used to modify the priors and soft reset parameters on-the-fly as well as estimating the position of the horizon in the images. The segmentation model can then be constrained by hard-assigning pixels above the horizon to the nonwater class. Temporal constraints on segmentation can be further imposed by image-based ego-motion estimation using techniques from structure-from-motion.

References

[1] H. Heidarsson and G. Sukhatme, "Obstacle detection from overhead imagery using self-supervised learning for autonomous surface vehicles," in *Proc. Int. Conf. Intell. Robots Syst.*, San Francisco, CA, USA, 2011, pp. 3160–3165.

[2] C. Rasmussen, Y. Lu, and M. K. Kocamaz, "Trail following with omnidirectional vision," in *Proc. Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, 2010, pp. 829–836.

[3] M. Montemerlo, S. Thrun, H. Dahlkamp, and D. Stavens, "Winning the DARPA grand challenge with an AI robot," in *Proc. AAAI Nat. Conf. Artif. Intell.*, Boston, MA, USA, 2006, pp. 17–20.

[4] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski, "Self-supervised monocular road detection in desert terrain," in *Proc. Robot. Sci. Syst. (RSS)*, Philadelphia, PA, USA, Aug. 2006. [Online]. Available: http://www.roboticsproceedings.org/rss02/index.html

[5] S. M. Ettinger, M. C. Nechyba, P. G. Ifju, and M. Waszak, "Vision-guided flight stability and control for micro air vehicles," *Adv. Robot.*, vol. 17, no. 7, pp. 617–640, 2003.

[6] Y. Lu and C. Rasmussen, "Simplified Markov random fields for efficient semantic labeling of 3D point clouds," in *Proc. Int. Conf. Intell. Robots Syst. (IROS)*, Vilamoura, Portugal, 2012, pp. 2690–2697.

[7] A. Rankin and L. Matthies, "Daytime water detection based on color variation," in *Proc. Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, 2010, pp. 215–221.

[8] S. Scherer *et al.*, "River mapping from a flying robot: State estimation, river detection, and obstacle mapping," *Auton. Robots*, vol. 33, nos. 1–2, pp. 189–214, 2012.

[9] C. Onunka and G. Bright, "Autonomous marine craft navigation: On the study of radar obstacle detection," in *Proc. Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Singapore, 2010, pp. 567–572.

[10] H. Heidarsson and G. Sukhatme, "Obstacle detection and avoidance for an autonomous surface vehicle using a profiling sonar," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Shanghai, China, 2011, pp. 731–736.

[11] L. Elkins, D. Sellers, and W. M. Reynolds, "The autonomous maritime navigation (AMN) project: Field tests, autonomous and cooperative behaviors, data fusion, sensors, and vehicles," *J. Field Robot.*, vol. 27, no. 6, pp. 790–818, 2010.

[12] T. H. Hong, C. Rasmussen, T. Chang, and M. Shneier, "Fusing ladar and color image information for mobile robot feature detection and tracking," in *Proc. Int. Conf. Intell. Auton. Syst.*, Marina del Rey, CA, USA, 2002, pp. 124–133.

[13] P. Santana, R. Mendica, and J. Barata, "Water detection with segmentation guided dynamic texture recognition," in *Proc. IEEE Int. Conf. Robot. Biomimet. (ROBIO)*, Guangzhou, China, 2012, pp. 1836–1841.

[14] D. Socek, D. Culibrk, O. Marques, H. Kalva, and B. Furht, "A hybrid color-based foreground object detection method for automated marine surveillance," in *Advanced Concepts for Intelligent Vision Systems*. Berlin, Germany: Springer, 2005, pp. 340–347.

[15] S. Fefilatyev and D. Goldof, "Detection and tracking of marine vehicles in video," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, USA, 2008, pp. 1–4.

[16] H. Wang *et al.*, "A vision-based obstacle detection system for unmanned surface vehicle," in *Proc. Int. Conf. Robot. Autom. Mechatronics*, Qingdao, China, 2011, pp. 364–369.

[17] T. Huntsberger, H. Aghazarian, A. Howard, and D. C. Trotz, "Stereo vision-based navigation for autonomous surface vessels," *J. Field Robot.*, vol. 28, no. 1, pp. 3–18, 2011.

[18] S. Khan and M. Shah, "Object based segmentation of video using color, motion and spatial information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Patt. Recognit.*, vol. 2. Kauai, HI, USA, 2001, pp. 746–751.

[19] T. M. Nguyen and Q. M. J. Wu, "A nonsymmetric mixture model for unsupervised image segmentation," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 751–765, Apr. 2013.

[20] N. Nasios and A. Bors, "Variational learning for Gaussian mixture models," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 849–862, Aug. 2006.

[21] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[22] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.

[23] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[24] P. F. Felzenszwalb and O. Veksler, "Tiered scene labeling with dynamic programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3097–3104.

[25] C. Wojek and B. Schiele, "A dynamic conditional random field model for joint labeling of object and scene classes," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France, 2008, pp. 733–747.

[26] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 2001, pp. 282–289.

[27] P. Kontschieder, S. Bulo, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, 2011, pp. 2190–2197.

[28] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 4, pp. 1907–1916, Aug. 2004.

[29] T. M. Nguyen and Q. Wu, "Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 193–202, Feb. 2012.

[30] S. Makrogiannis, G. Economou, S. Fotopoulos, and N. Bourbakis, "Segmentation of color images using multiscale clustering and graph theoretic region synthesis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 2, pp. 224–238, Mar. 2005.

[31] W. Tao, H. Jin, and Y. Zhang, "Color image segmentation based on mean shift and normalized cuts," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 5, pp. 1382–1389, Oct. 2007.

[32] M. Alpert, S. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, 2012, pp. 1–8.

[33] Z. Li, X. M. Wu, and S. F. Chang, "Segmentation using superpixels: A bipartite graph partitioning approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2012, pp. 789–796.

[34] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (ICCV)*, Nice, France, 2003, pp. 10–17.

[35] H. Lu, R. Zhang, S. Li, and X. Li, "Spectral segmentation via midlevel cues integrating geodesic and intensity," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2170–2178, Dec. 2013.

[36] A. Diplaros, N. Vlassis, and T. Gevers, "A spatially constrained generative model and an EM algorithm for image segmentation," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 798–808, May 2007.

[37] M. Kristan, J. Perš, V. Sulić, and S. Kovačič, "A graphical model for rapid obstacle image-map estimation from unmanned surface vehicles," in *Proc. Asian Conf. Comput. Vis.*, Singapore, 2014, pp. 1–67.

[38] M. Kristan, A. Leonardis, and D. Skočaj, "Multivariate online kernel density estimation with Gaussian kernels," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2630–2642, 2011.

[39] M. Kristan and A. Leonardis, "Online discriminative kernel density estimator with Gaussian kernels," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 355–365, Mar. 2014.

[40] S. Wang, J. Wang, and F. Chung, "Kernel density estimation, kernel methods, and fast learning in large data sets," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 1–20, Jan. 2014.

[41] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[42] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph. (SIGGRAPH)*, vol. 23, no. 3, pp. 309–314, 2004.

[43] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[44] S. Bagon. (Dec. 2006). *MATLAB Wrapper for Graph Cut*. [Online]. Available: http://www.wisdom.weizmann.ac.il/~bagon

[45] D. Tick, A. Satici, J. Shen, and N. Gans, "Tracking control of mobile robots localized via chained fusion of discrete and continuous epipolar geometry, IMU and odometry," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1237–1250, Aug. 2013.
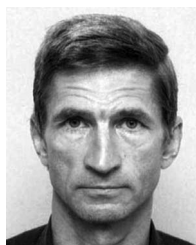
Matej Kristan (M'10) received the Ph.D. degree in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, in 2008.

He is currently an Assistant Professor with the Faculty of Computer and Information Science and the Faculty of Electrical Engineering, University of Ljubljana. He is a part of the organization board of the visual object tracking, visual object tracking, and challenges. His current research interests include probabilistic methods for computer vision and pattern recognition with focus on visual tracking, probabilistic dynamic models, online learning, object detection, and vision for mobile robotics.

Vildana Sulić Kenk received the B.Sc. and Ph.D. degrees from the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, in 2006 and 2011, respectively.

She is a member of the Machine Vision Laboratory, Faculty of Electrical Engineering, University of Ljubljana. Her current research interests include computer vision, machine vision, image processing, human-motion analysis, and visual-sensor networks.

Stanislav Kovačič (M'09) received the Ph.D. degree in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, in 1990.

He is a Full Professor and the Head of the Laboratory for Machine Vision, Faculty of Electrical Engineering, University of Ljubljana. From 1985 to 1988, he was a Visiting Researcher with General Robotics and Active Sensory Perception Laboratory, University of Pennsylvania, Philadelphia, PA, USA. He was also a Visiting Researcher and a Visiting Professor with the Technische Fakultat der Friedrich-Alexander Universitat, Erlangen, Germany, and with the Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia. His current research interests include various aspects of image and video analysis, image registration with applications in medicine, industry, and sports. He has authored/co-authored over 120 publications in journals, conferences, and book chapters.

Janez Perš (M'10) received the B.Sc., M.Sc., and the Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, in 1998, 2001, and 2004, respectively.

He is currently an Assistant with the Machine Vision Laboratory, Faculty of Electrical Engineering, University of Ljubljana. His current research interests include image sequence processing, object tracking, human motion analysis, dynamic motion-based biometry, and autonomous and distributed systems.