International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 2 (2015), pp. 89-104
http://dx.doi.org/10.14257/ijsip.2015.8.2.10

Single-camera Three-dimensional Tracking of Underwater Objects

Zhe Chen[1], Jie Shen[1], Tanghuai Fan[2], Zhen Sun[3] and Lizhong Xu[1]*

[1]*College of Computer and Information Engineering, Hohai University*
[2]*School of Information Engineering, Nanchang Institute of Technology*
[3]*College of Energy and Electrical Engineering, Hohai University*
chenzhe@hhu.edu.cn

Abstract

A novel method for tracking underwater objects in three-dimensional space is proposed. Specifically, by a single camera this method has the ability to estimate the three-dimensional trajectory of the interest objects. Mathematically the general framework of this method is formulated as a probabilistic problem. In addition to the two-dimensional location in the image plane, the parameter respect to the distance between objects and the viewer (depth) is introduced into the motion state descriptor. The haze concentration in the image is utilized as the cue to estimate the depth. In order to match the objects existing in successive frames, color models of the reference and candidates are related by special consideration on the dynamic appearance variability caused by the underwater environments and the motion of objects. As a result, the trajectory exactly describing the three-dimensional motion trend of the interest objects is produced by the video filmed by a single camera. We present results over real underwater videos. Both single-object and multi-object tracking in three-dimensional space are achieved by using the proposed method.

Keywords: *Three-dimensional tracking, Particle filter, Depth information estimation*

1. Introduction

The increasing interest in the investigation of the motion of underwater objects has heightened the need for developing tools that provide robust quantitative data [1, 2]. The underwater computer vision has emerged as an attractive technology for motion analysis; however, there are still limited ways to monitor the underwater object motion in a quantitative manner. Due to the specially challenging environments, the underwater videos are seriously distorted in terms of the clarity and color fidelity, so that in comparison to the ground-based tasks the interest objects in the underwater video is more difficult to be extracted [3, 4]. Moreover, the tracking tasks are further complicated by the large motion freedom of the underwater objects, which would result in the abrupt change of the appearance and the moving trajectory [5]. At last, the wide use of moving cameras poses a new challenge. The entire image positional array of the video is time-varied, distorting the motion vector of the objects. Hence the strategy based on the fixed camera is no longer available and the trajectory extracted has to be corrected by introducing the motion vector of the camera [6, 7].

In this paper, a systematic method for tracking underwater object motion in three dimensions by a single camera is proposed. First, we formulate the tracking as a probabilistic problem. The parameter of the depth is injected into the descriptor of object states, forming the framework of three-dimensional tracking. Second, we introduce the dark channel prior based method for estimating the haze concentration which further underlies the depth estimation of objects. Finally, the image is restored and the refined

ISSN: 2005-4254 IJSIP

color model of the underwater objects is estimated. Our primary innovation is the use of a single camera for markerless, real-time tracking of single or multiple underwater objects in the three-dimensional space.

2. Tracking in Three Dimensional Space

To the best of our knowledge, no existing technologies can use a single camera to track the interest object in three dimensions, to say nothing of three-dimensional tracking underwater objects. Commonly, the three-dimensional structure of object is calculated by stereoscopic vision systems [8, 9]. In these processes, the object and the camera should be correctly calibrated to construct the relation between the image coordinate system and the world coordinate system [10]. This limitation obviously cannot cater for the need for underwater object tracking tasks, since most of the underwater monitor systems take the video filmed by a single camera as the input. Hence, aimed to efficiently estimate the three-dimensional trajectory of the underwater object motion, we require a more feasible method which has the ability to estimate the object depth by a single frame.

Inspired by the work of He [11], it is noted that the haze concentration in the image can be taken as the cue to estimate the depth. The dark channel prior based haze estimation provides us a light to locate the underwater objects respected to the depth. The proposed method is formulated as a framework of probabilistic tracking problem. The solution is generally given by three key points.

I. Depth estimation: This step extracts the haze concentration in each frame, followed by reconstructing the depth of objects. Here, the dark channel prior and underwater optical model are combined.

II. Refinement of the object model: This step is achieved by inversely transforming the underwater optical model. The haze effect is removed and the color distortion of the object appearance is compensated successively. Further, based on the depth estimation, the initial window of the candidates is predetermined. Finally, the refined color models are established by statistical treatment on the color information in the window.

III. Compensation of cameral motion: Taking the motion of the camera into consideration, this step is to estimate the image motion parameter for correcting the object trajectory. The interest points in each frame are extracted, followed by the point matching. The SURF feature is taken in this process.

2.1. The Framework of Tracking Underwater Objects

For the high motion freedom of the underwater objects in water, the familiar problems for object tracking, such as abrupt variability of the appearance, motion state and the partial or full object occlusions [5], would arise more frequently. In this case, the deterministic methods, such as the MEG tracker [12] and GOA tracker [13], are weak for their rigid disciplines. However, the probabilistic methods, such as Kalman filter [14] and Particle filter [15], are the attractive options, since they are more robust against the noise, appearance variability and information loss . Hence, we formulate the underwater tracking task as a probabilistic problem, specialized to the particle filter which is capable of estimating the evolution of nonlinear, non-Gaussian stochastic processes.

2.1.1. Descriptor of the Dynamic State of Underwater Objects

Different from the tracking problem in the image plane that merely locating the objects by a pair of coordinates (O_x, O_y), in three-dimensional tasks the location of the objects is described by a state vector comprised of three global parameters.

$$\mathrm{x}_t = (O_x, O_y, O_z) \qquad (1)$$

Where, O_x, O_y and O_z are the coordinates of the object's origin respect to the axis of

x, y and z respectively. The task of object tracking is equivalent to the estimation of the parameters x_t at discrete time points. The transformation between states can be described by a system dynamics model as the sequence evolves.

$$\mathrm{x}_t = f_t(\mathrm{x}_{t-1}, v_{t-1}) \qquad (2)$$

where f_t is a function respect to the previous state, which describes the linear or nonlinear system dynamics, and v_{t-1} is an i.i.d.noise sequence. Measurements are taken from a noisy observation model respect to the states and the noise at the time t.

$$\mathrm{z}_t = h_t(\mathrm{x}_t, n_t) \qquad (3)$$

where h_t is a possibly linear or nonlinear function respect to the current state, and n_t is another i.i.d. noise sequence.

In order to formulate the object tracking as a stochastic state estimation problem, the first-order Markov chain is used to model the dynamic system. Accordingly, the equation (2) can be modeled by a prior probability density function, as $p(\mathrm{x}_t|\mathrm{x}_{t-1})$ and the equation (3) is transformed into a likelihood probability density function, as $p(\mathrm{x}_t|\mathrm{z}_t)$. Given the initial posterior $p(\mathrm{x}_0|\mathrm{z}_0) \equiv p(\mathrm{x}_0)$ and a series of independent observations from the initial point to the time t, as $\mathrm{z}_{1\cdots t}$, the probability of belief that the system is in state x_t can be evaluated by a posterior probability density function $p(\mathrm{x}_t|\mathrm{z}_{1\cdots t})$. From a probabilistic perspective, the tracking is the process to seek the optimal Bayesian solution to the poster probability density function frame by frame.

2.1.2. Particle Filtering

Particle filtering generally falls into the category of the Sequential Monte Carlo method which implements a recursive process to optimally estimate the posterior density $p(\mathrm{x}_t|\mathrm{z}_{1\cdots t})$. The key idea of particle filter is to represent the posterior density of any state by a set of samples $\left\{\mathrm{x}_t^i\right\}_{i=1}^N$ with associated weights $\left\{w_t^i\right\}_{i=1}^N$ [15]. The combination $\left\{\mathrm{x}_t^i, w_t^i\right\}_{i=1}^N$ is the random measurements which are call of the particles. Commonly, the weights are normalized as $\sum_{i=1}^N w_t^i = 1$. Then the posterior density in the time t can be approximated by the particle mass in a discrete weighted form, as follows:

$$p(\mathrm{x}_t|\mathrm{z}_{1:t}) \approx \sum_{i=1}^N w_t^i \delta(\mathrm{x}_{0:t} - \mathrm{x}_{0:t}^i) \qquad (4)$$

The weight of the particles is generated by the sequence importance sampling

$$w_t^i \propto p(z_t \mid x_t^i) \qquad (5)$$

To reduce the effects of particle degeneracy, the re-sampling should be performed one very filter iteration. As a result, majority of the particles fall around the peaks of the posterior and few computational resources are assigned on the margin. Thus, the state x_t can be approximated by the weighted samples, as follows:

$$\mathrm{x}_t = \sum_{i=1}^N w_t^i x_t^i \qquad (6)$$

2.2. Estimation of the Depth of the Underwater Objects

Generally, two strategies have the ability to extract the depth feature in underwater images. The stereo vision is the most popular technology. Their superiority lies in the ability to produce the precise in-scale measurement. However, the high hardware cost for establishing the binocular vision and high computational cost for rigidly calibrating the cameras are the two major bottlenecks blocking the application in practice. Alternatively, the depth estimation is also achieved by the monocular vision systems. Generally, in order to estimate the depth, any stringent but powerful prior should be introduced into the calculation, such as the dark channel prior [11] and the contrast constrains [16]. As only one single image is required, these methods are characterized by their high feasibility and generalization in contrast to the stereovision systems. However, due to lack of the calibration process, the estimation result is not metric but up-to-scale. Note the purpose of object tracking is to describe the trajectory reflecting the motion trend in previous frames, so that the up-to-scale estimation can still cater for the need of describing the motion in depth. Moreover, the relatively economical cost in hardware gives these single camera based methods more opportunities to be the key for the on-line applications.

The dark channel prior based depth estimation [11] is preferred in this paper. As water is a typical scattering medium, we can generalize the hazy image formation model to simulate the light propagation in water, as follows:

$$I_\lambda(\mathrm{x}) = J_\lambda(\mathrm{x})\rho_\lambda(\mathrm{x})\exp[-\alpha_\lambda r(\mathrm{x})] + B_\lambda(1-\exp[-\alpha_\lambda r(\mathrm{x})]) \quad (7)$$

Where, λ is the color channel, $I_\lambda(\mathrm{x})$ is the acquired photons at point x, $J_\lambda(\mathrm{x})$ is ambient light, $\rho_\lambda(\mathrm{x})$ is the reflectivity, α_λ is the wave-selective attenuation factor, $r(\mathrm{x})$ is the depth and the background light in water is represented as B_λ. The dark channel prior is discovered based on the statistics that if the image is haze-free, in most of the non-background patches at least one color channel has very low intensity at some pixels, $I_{dark}(\mathbf{x}) = \min_{\lambda\in\{r,g,b\}}(\min_{\mathbf{y}\in\Omega_x} I_\lambda(\mathbf{y}))$.The low intensities in the dark channel prior are mainly caused by the dark, colorful objects or the shadow. These three factors can be roughly generalized as a much low reflectively in any channel, $\rho_{dark}(\mathbf{x}) = \min_{\lambda\in\{r,g,b\}}(\min_{\mathbf{y}\in\Omega_x} \rho_\lambda(\mathbf{y})) \approx 0$. Hence, the dark channel is the representation to the haze concentration.

$$\begin{aligned}\min_{\lambda\in\{r,g,b\}}(\min_{\mathbf{y}\in\Omega_x} I_\lambda(\mathbf{y})) &= J_{dark}(\mathbf{x})\rho_{dark}(\mathbf{x})\exp[-\alpha_{dark}(\mathbf{x})r(\mathbf{x})] + B_{dark}(1-\exp[-\alpha_{dark}r(\mathbf{x})]) \\ &\approx B_{dark}(1-\exp[-\alpha_{dark}r(\mathbf{x})])\end{aligned} \quad (8)$$

When the background light is homogenous, the depth $r(\mathrm{x})$ can be calculated as:

$$r(\mathbf{x}) = -\ln((B_{dark} - I_{dark}(\mathbf{x}))/B_{dark})/\alpha_{dark} \quad (9)$$

In this solution, there are two parameters remaining unknown: B_{dark} and α_{dark}. For the researcher in the field of the computer vision, the attenuation factor α_{dark} is commonly given in Table 1. The background light in the dark channel B_{dark} can be derived from equation (9), as follows:

$$\max_{\mathrm{x}}(\min_{\lambda\in\{r,g,b\}}(\min_{\mathrm{y}\in\Omega_x} I_\lambda(\mathbf{y}))) = B_{dark}(1-\max_{\mathrm{x}}(\exp[-\alpha_{dark}r(\mathbf{x})])) \quad (10)$$

If there is any background area where the depth is numerically infinite, the attenuation of these points is equal to 0, as follows:

$$\max_{\mathrm{x}}(\exp[-\alpha_{dark}r(\mathbf{x})] = 0 \quad (11)$$

[1] We selected the type II water as the reference as $\alpha_{red} = 0.825, \alpha_{green} = 0.95, \alpha_{red} = 0.97$ [3] N. G. Jerlov, *Optical oceanography* vol. 5: Access Online via Elsevier, 2009.

The background light in the dark channel B_{dark} is given as follows:

$$B_{dark} = \max_{\mathrm{x}}(\min_{\lambda\in\{r,g,b\}}(\min_{\mathrm{y}\in\Omega_{\mathrm{x}}} I_{\lambda}(\mathbf{y}))) \quad (12)$$

Finally, the depth of any point can be derived by combining equation (12) and equation (9), as follows:

$$r(\mathbf{x}) = -\ln\left(\frac{\max_{\mathrm{x}}(\min_{\lambda\in\{r,g,b\}}(\min_{\mathrm{y}\in\Omega_{\mathrm{x}}} I_{\lambda}(\mathbf{y}))) - \min_{\lambda\in\{r,g,b\}}(\min_{\mathrm{y}\in\Omega_{\mathrm{x}}} I_{\lambda}(\mathbf{y}))}{\max_{\mathrm{x}}(\min_{\lambda\in\{r,g,b\}}(\min_{\mathrm{y}\in\Omega_{\mathrm{x}}} I_{\lambda}(\mathbf{y})))}\right) \Big/ \alpha_{dark} \quad (13)$$

Figure 1 shows the result of the depth estimation given by the dark channel prior. We can find that the intensity gradient is in correspondence to the depth variation of points, as the nearest turtle is pigmented with the black, while the coral and rocks with the gray and the background is corresponding to the brightest region in the depth map.

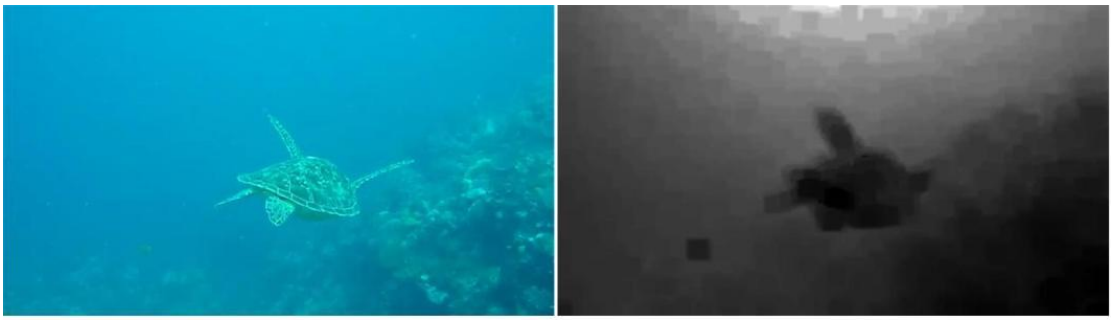

Figure 1. Depth Estimation, (a) Underwater Frame (b) Depth Map Estimated in the 8×8 Windows

2.3 Refined Object Model

Since the gravity is partly or completely counteracted by the buoyancy of water, the motion of underwater objects might freely change and their pose would be dramatically variable as well. Consequently, the space-dependent appearance, which is the projection of the three-dimensional structure into the image plane, does not have the robust ability to exactly model objects and would possibly break down in any frame. To solve this problem, the color based model [17] is especially appealing.

In this paper, the histogramming technique is employed to establish object color model as follows: first, the chromatic information is decoupled in the Hue-Saturation-Value (HSV) space. Second, the HS histogram with $L_h L_s$ bins is produced by statistical treatment on the pixels with saturation and value larger than the thresholds (which are set to 0.08 and 0.18 in the experiments). The remaining pixels are used to produce L_v additional value-only bins. As the whole, the complete color histogram model includes $L = L_h L_s + L_v$ bins. For any single pixel $\mathbf{x}$ in frame t, the histogram color model $\mathbf{M}_t(\mathbf{x})$ can be established as follows:

$$\mathbf{M}_t(\mathbf{x}) = \left(M_t^1(\mathbf{x}), M_t^2(\mathbf{x}), \cdots, M_t^L(\mathbf{x})\right) \quad (14)$$

Where $M_t^{1\cdots L}(\mathbf{x})$ is the statistical information in the corresponding bins. However, in spite of the super performance, both the depth-varied haze and the wave-selective attenuation in the water medium would seriously decay the color models. Moreover, without any prior, the initial candidate region where the color information is extracted has to be blindly controlled by a window with a certain size.

To refine the color model, two technologies are successively operated on the frame. On the one hand, the adaptive window is proposed based on the camera projection model and takes the depth information as the key to determine the initial size. On the other hand, color restoration is operated based on the underwater optical model.

Figure 2. Adaptive Window for a Single Object (a) Nearby One Enveloped by Larger Window (b) Faraway One Enveloped by Smaller Window

2.3.1 Adaptive Window

According to the computer vision geometry, the image formation process drops the object structure from three-dimensional world to a two-dimensional image plane. However, the depth also has strong relation to the representation of the object, especially in the underwater environments. As we know, the faraway objects commonly correspond to a smaller region, while larger region is filled with nearby ones (Figure 2). To a single object, whose projection size, ignoring the changes in gesture, would be inverse to the depth. Hence we can actively set the candidate window by considering the depth variability of objects between frames; especially we relate the window size between frames, as follows

$$R_t^i = k\frac{r_{t-1}}{r_t^i}R_{t-1} \qquad i = 1,\cdots,N \qquad (15)$$

Where, R_t^i is the i th candidate object regions in the current frame t, R_{t-1} is the window of the object region in previous frame $t-1$, r_t^i, r_{t-1} are the depth of the origin of the i th candidate in the current frame t and $t-1$, and the moderation is controlled by factor k. However, once the object gesture abruptly changes, the projection size of objects would change accordingly. In this case, the initial window can not tightly envelope the objects. To solve this problem, a relaxer is introduced as:

$$R_t^i + \xi_t^i \qquad (16)$$

Where, ξ_t is the relax factor describing the size moderation caused by the gesture variability.

2.3.2 Color Restoration

According to the image formation model, underwater object appearance is seriously degraded by the haze and color-selective attenuation. Theoretically, these two effects are directly determined by the distance between the points and the camera. Given the depth, the object color can be restored by inversely transforming the underwater optical model, as follows:

$$I_\lambda^O = J_\lambda(\mathrm{x})\rho_\lambda(\mathrm{x}) = \left(I_\lambda(\mathrm{x}) - B_\lambda(1-\exp[-\alpha_\lambda r(\mathrm{x})])\right)/\exp[-\alpha_\lambda r(\mathrm{x})] \quad \lambda \in \{r,g,b\} \tag{17}$$

(a) (b)

Figure 3. Object Color Restoration (a) Degraded Color (b) Restored Color

Figure 3 shows the degraded color in the raw image (Figure 3(a)) and result restored by the Equation 17 (Figure 3(b)). By the visual inspection, it is clear that the color fidelity is enhanced and the haze noise is removed as well. With the restored image, the model establishment can be achieved by a kernel function, as:

$$q_t^i(l,\mathbf{x}) = \eta \sum_{\mathbf{X}\in R_t^i} w(\left|\mathbf{x}-\mathbf{u}_t^i\right|)\delta(b_t(\mathbf{u}_t^i)-l) \quad l=1,\cdots,L \tag{18}$$

Where, $q_t^i(l,\mathbf{x})$ is the l th bin of the histogram at point $\mathbf{x}$ belonged to the i th candidate at the frame t, $\mathbf{u}_t^i$ is the origin of the candidate region, w is the weight, η is a normalization constant ensuring $\sum_{l=1}^{L} q_t^i(l,\mathbf{x}) = 1$, $b_t(\mathbf{u}_t^i)$ is the bin index of the color vector in the centroid. As mentioned above, the shape of the underwater objects might dramatically change continually in a video sequence, the reference distribution hence is gathered rather than at the initial time but at a previous one time q_{t-1}. The similarity between the reference and candidates is measured by the Bhattacharyya similarity coefficient, as follows:

$$D(q_{t-1}, q_t^i(\mathbf{x})) = \left[1 - \sum_{l=1}^{L} \sqrt{q_{t-1}(l) q_t^i(l,\mathbf{x})}\right]^{\frac{1}{2}} \tag{19}$$

The probability $p(z_t \mid \mathbf{x}_t) \propto p(D^2(q_{t-1}, q_t^i(\mathbf{x})))$ is exponentially monotonic to the distance D^2, as follows:

$$p(z_t \mid \mathbf{x}_t) \propto \exp\left(-\gamma D^2(q_{t-1}, q_t^i(\mathbf{x}))\right) \tag{20}$$

We set the parameter γ as $\gamma = 20$ in the experiments.

2.4 Cameral Motion Compensation

Differing from the ground-based systems which usually take the fixed cameras as the sensors, the underwater videos are commonly recorded by the moving cameras which are handed by the diver or loaded in the automatic underwater vehicle. This makes a novel challenge to the object tracking that we have to correct the trajectory by compensating the camera ego motion.

The camera motion estimation can be divided into three main steps: the detection of interest point; description of interest point and the correspondence to interest points. For the super performance and the low computational cost, the SURF feature is introduced in this phase.

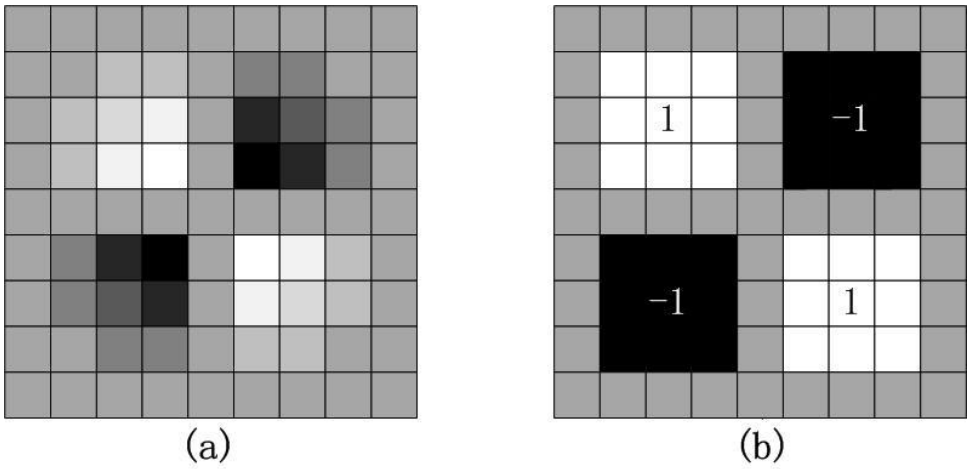

Figure 4. (a) Gaussian Second Order Derivative in x, y-direction. (b) Corresponding Box filter Approximation

2.4.1 Interest Point Detection

The detector of the SURF is based on the Hessian matrix which is more stable and repeatable than the counterparts [18]. Given a point $\mathbf{x}$ in the restored image I^O, the Hessian matrix $H(\mathbf{x},\sigma)$ in $\mathbf{x}$ at scale σ is defined as follows:

$$H(\mathbf{x},\sigma)=\begin{bmatrix} L_{xx}(\mathbf{x},\sigma)\, L_{xy}(\mathbf{x},\sigma) \\ L_{xy}(\mathbf{x},\sigma)\, L_{yy}(\mathbf{x},\sigma) \end{bmatrix} \quad (21)$$

Where $L_{xx}(\mathbf{x},\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I^O, similar to the other three elements. Inspired by the Low's success with the Laplacian of Gaussian (LoG) approximations, the second order Gaussian derivatives is further approximated with the box filter which is call of the Difference of Gaussian (DoG) (Figure 4) .

The scale starts with the 9×9 window, then it is extended to the size 15×15, 21×21 and 27×27. The location of the interest point in the image and over scales relies on the determinant of the Hessian matrix. Non-maximum suppression in a $3\times 3\times 3$ neighborhood is applied and the maxima of the determinant of the Hessian matrix are then interpolated in scale and image space by the method proposed by Brown and Lowe.

2.4.2 Interest Point Description

The distribution of intensity content within the neighborhood of the interest point is taken as the descriptor of the interest point. The gradient information extracted by the SIFT is updated by the first order Haar wavelet responses in x and y directions. First, we construct a square region centered on the interest point and oriented along the selected orientation. Then, the interest point neighborhood region is regularity split into smaller 4×4 square sub-regions where the Haar wavelet responds is computed. Finally, the wavelet responds in the horizontal and vertical directions are summed up over each sub-region, forming the first two elements in the descriptor vector. The other two elements are extracted by the sum of the absolute value of the responds. On the whole, the descriptor is expressed as follows:

$$\mathbf{v}=\left(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right) \quad (22)$$

Where, d_x and d_y are the Haar wavelet responses in x and y directions.

Figure 5. SURF Feature based Frame Correspondence

2.4.3 Interest Point Correspondence and Cameral Motion Calculation

In the SURF feature based framework, the sign of the Laplacian is employed to accelerate the process of point correspondence. Only points with same type of contrast are mutually compared. Hence, this minimal information allows for faster corresponding and gives a slight increase performance (Figure 5). Besides the in-plane translation of the camera, since the depth is given for each interest point, the three dimensional translation of the camera can be calculated, as:

$$\mathbf{MV}(t) = \left(p_x^r(t) - p_x^c(t-1), p_y^r(t) - p_y^c(t-1), p_z^r(t) - p_z^c(t-1) \right) \quad (23)$$

Where, $\left(p_x^r(t), p_y^r(t), p_z^r(t) \right)$ and $\left(p_x^c(t-1), p_y^c(t-1), p_z^c(t-1) \right)$ are the location of the reference interest point in the frame t and $t-1$. With the motion vector of camera, the location of the object in the frame t is corrected as follows:

$$\mathbf{x}_t' = \mathbf{x}_t - \mathbf{MV}(t) \quad (24)$$

3. Experimental Results

Here, we present the results in two typical contexts with single or multiple objects. The video is downloaded from Youtube [17-21]. Single object tracking experiments are firstly presented to prove the availability of the proposed method for the three-dimensional tracking, including the tasks which require camera motion compensation or not. Then the proposed method is taken to solve the multiple-object tracking problem and the performance is also evaluated in the complicated situation where the camera drastically shakes.

3.1. Single-object Tracking

The selected sea turtle sequence (Figure 1(a)) has 20 frames of 1280×720 pixels, and the movement of the turtle is tracked. This object is initialized with a hand-drawn square region. The particle filter based tracker proves to be robust to the clutter and distracters. The restored color information of the object provides a wealth of available features (Figure 3(b)). By the dark channel prior based estimation, the depth information is given (Figure 1(b)). Since the camera is almost static in this short sequence, the motion vector of the camera is set as $\mathbf{MV}_t = (0,0,0)$.Consequently, the three-dimensional trajectory describing the motion trend of the sea turtle is estimated (Figure 6). Due to lack of the ground truth data, we can hardly quantify the correctness of the trajectory. However, through visual evaluation, we also have full confidence to confirm that the estimated trajectory has the ability to describe the motion trend of the sea turtle at least.

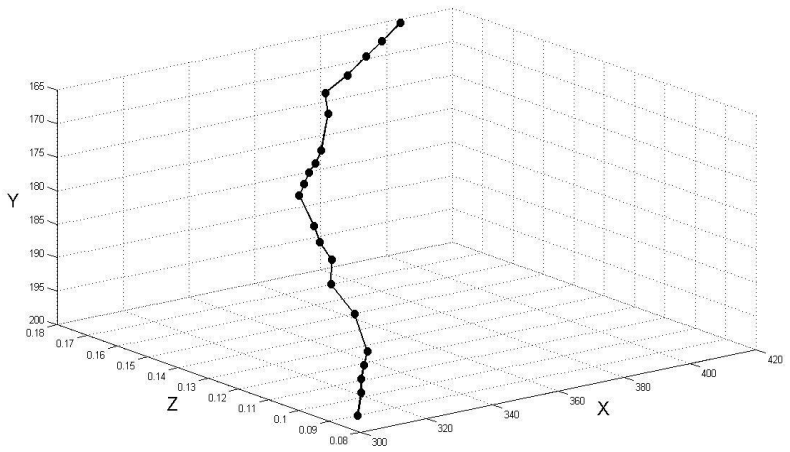

Figure 6. Three-dimensional Trajectory of the Sea Turtle Filmed by a Static Camera

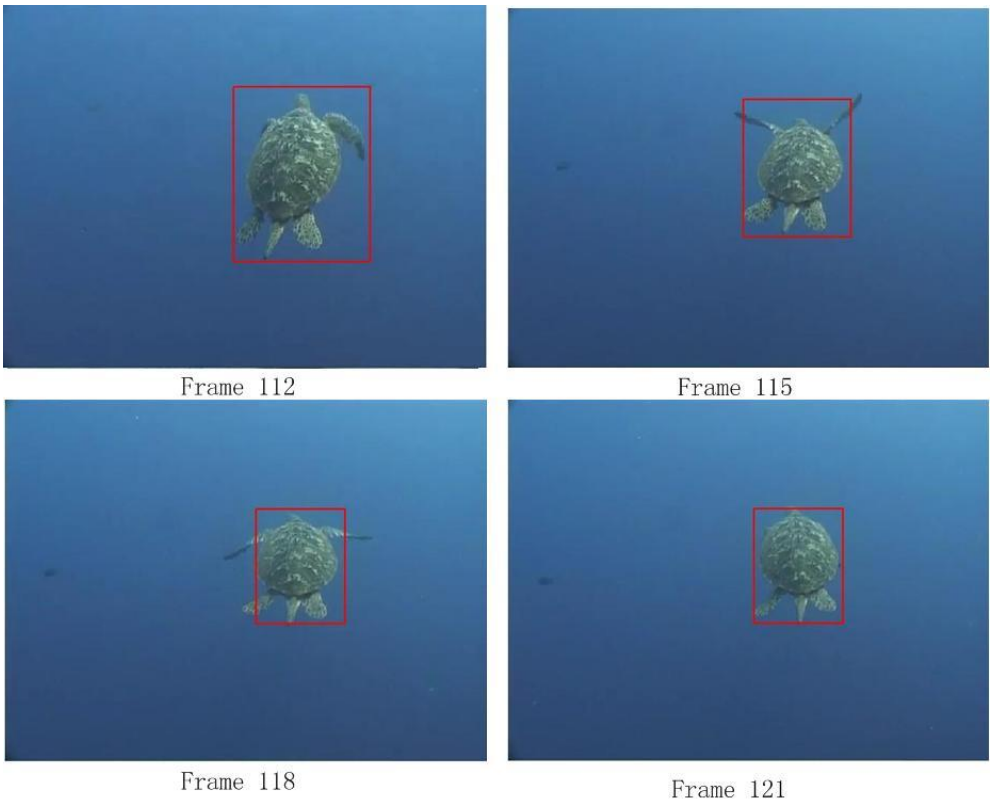

Figure 7. Single-sea Turtle Tracking in the Video Acquired by a Fixed Camera

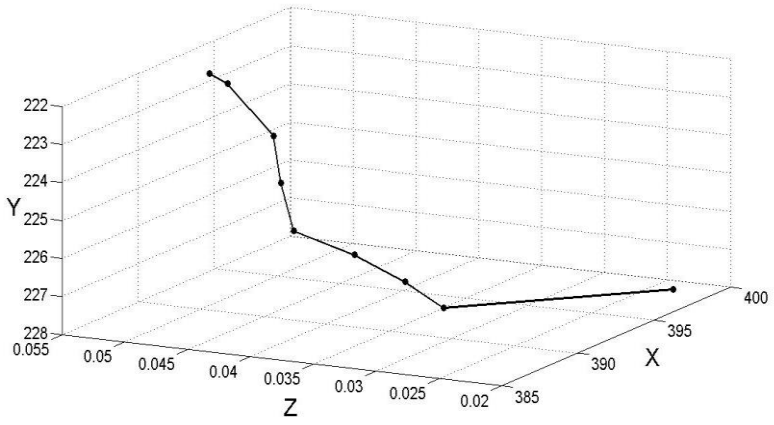

Figure 8. Three-dimensional Trajectory Corresponding to the Video of Figure 7

One more result is achieved on another sea turtle sequence with fixed camera (Figure 7). Figure 8 shows the trajectory estimated, which also exactly mirror the motion trend of the interest sea turtle.

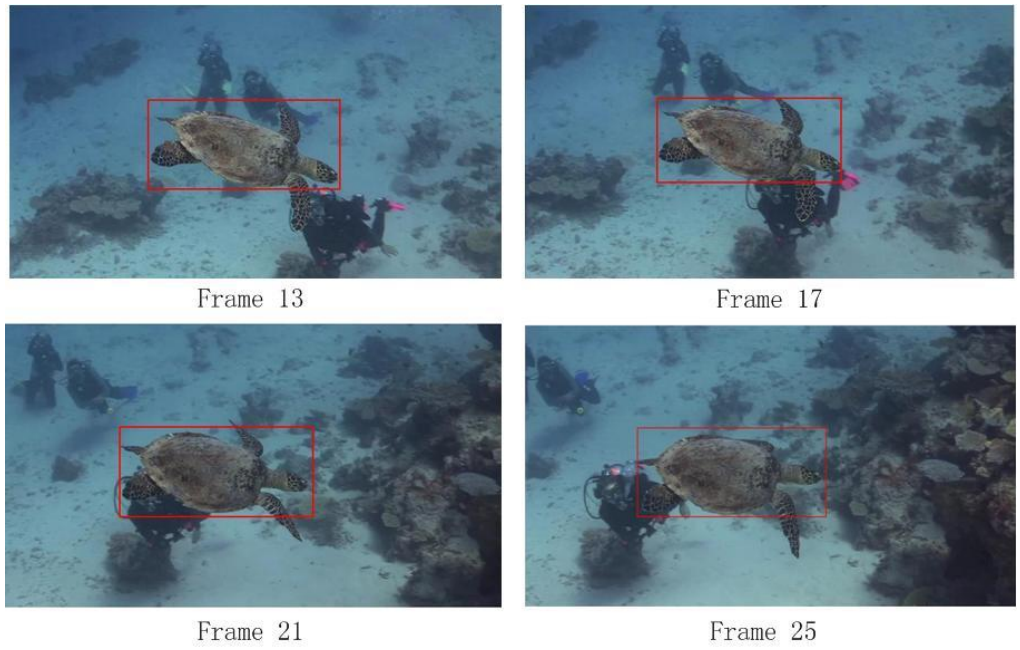

Figure 9. Single Sea Turtle Tracking in the Video Filmed by the Moving Camera

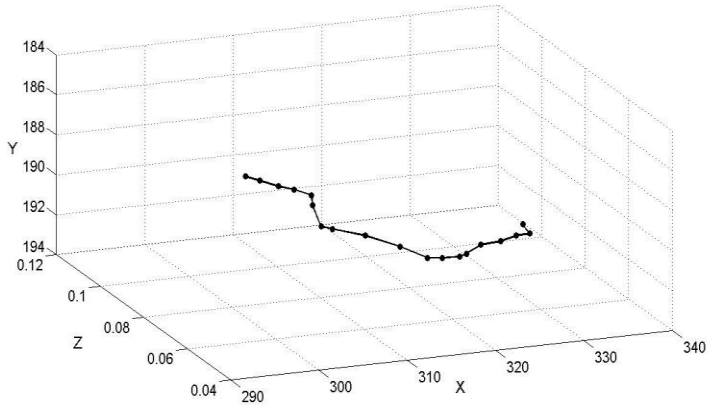

Figure 10. Three-dimensional Trajectory of the Sea Turtle Filmed by a Moving Camera

Given the video filmed by the moving camera, it is necessary to perform the motion compensation process. The selected sequence (Figure 9) has 19 frames with size of 1280×720 pixels, and the movement of the third sea turtle is tracked. In the phase of SURF correspondence, the interest points located in the regions of the reference or the candidate are abandoned and 40 interest points left (Figure 5). As the result, Figure 10 shows the estimated trajectory. We can find that our method is robust against the influence of the cameral motion and the essential trajectory of the sea turtle motion is given.

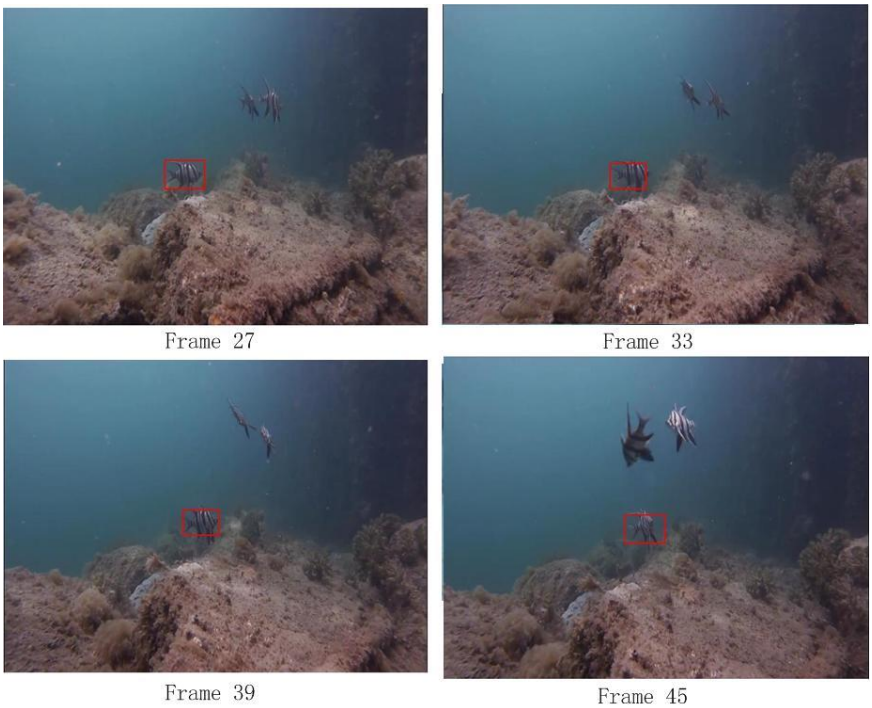

Figure 11. Single-fish Tracking in the Video Filmed by the Moving Camera

Figure 12. SURF Feature based Frame Correspondence

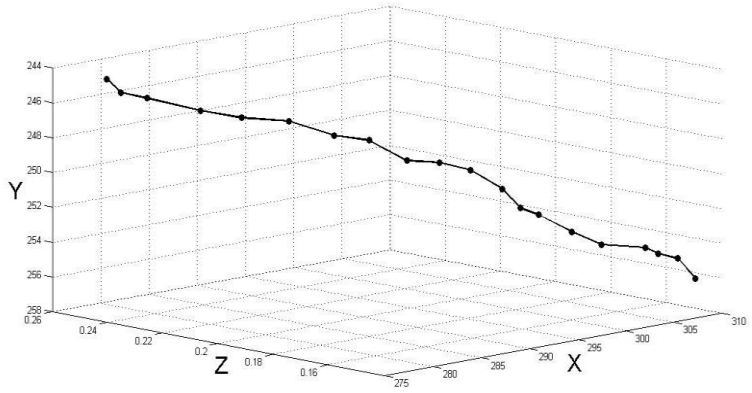

Figure 13. Three-dimensional Trajectory of a Fish Filmed by a Moving Camera

The last experiment of single object tracking is performed on a fish video filmed by a moving camera. One of the fish is tracked in 20 frames with size of 1280×720 pixels (Figure 11). Two similar fish swimming around the interested object form the strong distracter to the tracker. Hence, this situation is more complicated than the former one. 20 interest points are selected for image correspondence (Figure 12). The result proves that the particle filter based three-dimensional tracking can well adapt to this rigorous situation and still has the ability to produce a high-quality trajectory reflecting the motion trend of the interest fish (Figure 13).

3.2. Multi-object Tracking

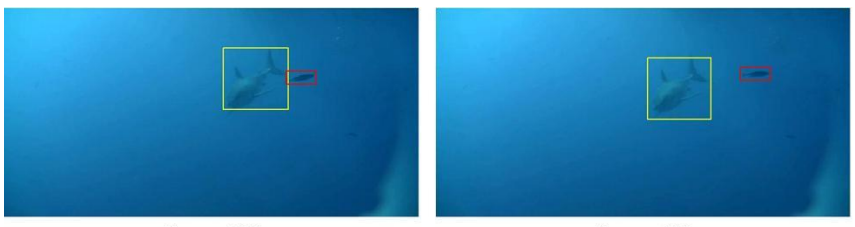

Figure 14. Multi-fish Tracking in the Video Filmed by the Static Camera

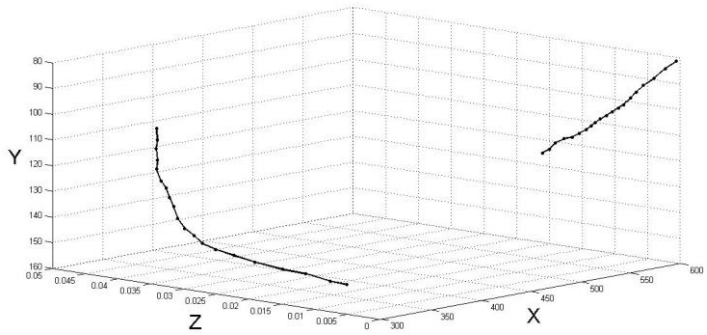

Figure 15. Three-dimensional Trajectories of the Shark and the Little Fish Filmed by a Static Camera

The proposed method is further evaluated by the multi-object tracking experiments. The selected sequence (Figure 14) has 20 frames with size of 1280×720 pixels, and the motion of a shark and the little fish are simultaneously tracked. As the camera is almost static in this short sequence, the motion vector of the camera is set as $\mathbf{MV} = (0,0,0)$.Consequently, the three-dimensional trajectories describing the motion trend of the shark and the little fish are both simulated in a uniform coordinate system (Figure 15). By visual evaluation, the trajectories are correctly in line with the motion trend of objects.

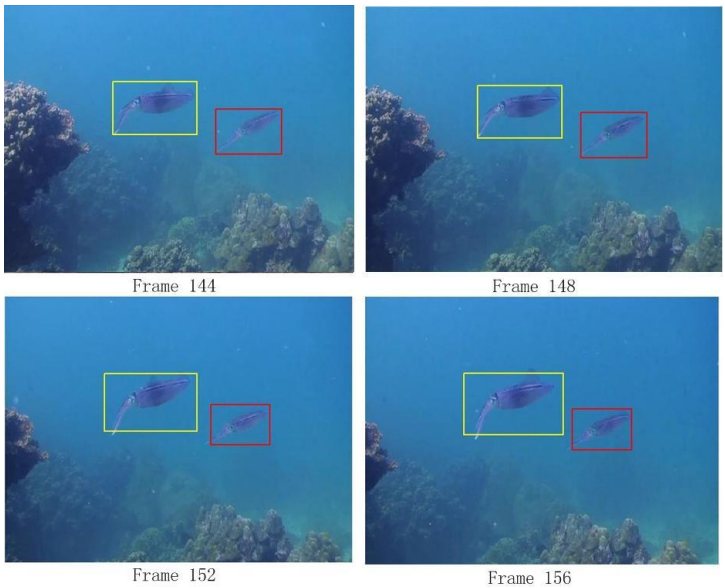

Figure 16. Multiple Sleeve-fish Tracking in the Video Filmed by the Moving Camera

Figure 17. SURF Feature based Frame Correspondence

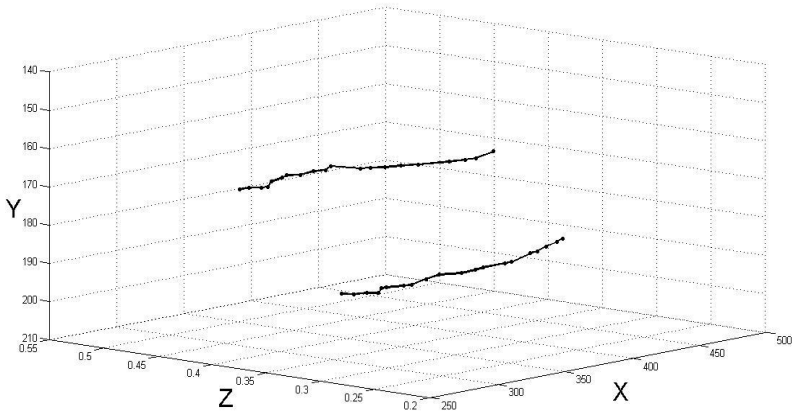

Figure 18. Three-dimensional Trajectory of Tow Sleeve-fish Filmed by a Moving Camera

One more result of multi-object tracking is operated on the tow sleeve-fish video filmed by a moving camera. The selected sequence (Figure 16) has 19 frames with size of 1280×720 pixels. 10 interest points are selected for image correspondence (Figure 17). As a result, the movement trend of two sleeve-fish is correctly represented by a pair of three-dimensional trajectories (Figure 18).

4. Conclusion

In this paper, we have proposed the method for tracking underwater objects in three-dimensional space. The general tracking framework is formed as the classical particle filter, and the motion states of the objects are described in the three dimensional space. Aimed to estimate the depth of the interest objects in the underwater video, the dark channel prior is introduced. Moreover, taking the camera motion into consideration, the SURF feature based image correspondence is employed to calculate the three-dimensional translation of the camera. Finally, the trajectory is corrected by cameral motion compensation. To our best knowledge, this is a novel work which uses a single-camera to achieve three-dimensional object tracking. However, note that the dark channel prior based estimation can only provide us with an up-to-scale depth estimation. Hence, the trajectory estimation in this paper is not the precise measurement but the reflection to the motion trend of the underwater objects.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No.61263029)

References

[1]. P. Domenici and R. Blake, "The kinematics and performance of fish fast-start swimming", Journal of Experimental Biology, vol. 200, **(1997)**.
[2]. S. Butail and D. A. Paley, "Three-dimensional reconstruction of the fast-start swimming kinematics of densely schooling fish", Journal of The Royal Society Interface, vol. 9, **(2012)**.
[3]. N. G. Jerlov, "Optical oceanography", Elsevier, **(2009)**.
[4]. G. Telem and S. Filin, "Photogrammetric modeling of underwater environments", ISPRS Journal of Photogrammetry and Remote Sensing, vol. 65, **(2010)**.
[5]. A. Yilmaz, O. Javed and M. Shah, "Object tracking: A survey", Acm Computing Surveys (CSUR), vol. 38, **(2006)**.
[6]. P. Burt, J. Bergen, R. Hingorani, R. Kolczynski, W. Lee, A. Leung, J. Lubin and H. Shvayster, "Object tracking with a moving camera", Proceedings of workshop on Visual Motion, **(1989)** March 20-22, UK.
[7]. K. Jinman, I. Cohen and G. Medioni, "Continuous tracking within and across camera streams", Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **(2003)** June 19-20, UK.
[8]. S. -W. Min, B. Javidi and B. Lee, "Enhanced three-dimensional integral imaging system by use of double display devices Applied Optics", vol. 42, **(2003)**, USA.
[9]. A. D. Straw, K. Branson, T. R. Neumann and M. H. Dickinson, "Multi-camera real-time three-dimensional tracking of multiple flying objects", Journal of The Royal Society Interface, vol. 8, **(2011)**.
[10]. P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes", Robotics, vol. 3 **(1992)**.
[11]. K. He, J. Sun, X. Tang, "Single image haze removal using dark channel prior", IEEE Conference on Computer Vision and Pattern Recognition, vol. 4, **(2009)** June 20-25, USA.
[12]. V. Salari and I. K. Sethi, "Feature point correspondence in the presence of occlusion", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, **(1990)**.
[13]. C. J. Veenman, M. J. T. Reinders and E. Backer, "Resolving motion correspondence for densely moving points", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, **(2001)**.
[14]. T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, **(1986)**.
[15]. A. Doucet, S. Godsill and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering", Statistics and computing, vol. 10, **(2000)**.
[16]. Z. Chen, H. Wang, J. Shen, X. Li and L. Xu, "Region-Specialized Underwater Image Restoration in Inhomogeneous Optical Environments", Optik-International Journal for Light and Electron Optics, vol. 9, **(2014)**.
[17]. P. Pérez, C. Hue, J. Vermaak and M. Gangnet, "Color-based probabilistic tracking", Europeon Conference on Computer Vision, **(2002)**, pp. 661-675.
[18]. H. Bay, A. Ess and T. Tuytelaars, "Speeded-up robust features (SURF)", Computer Vision and Image Understanding, vol. 110, **(2008)**.

