International Journal of Cognitive Informatics and Natural Intelligence
Volume 14 • Issue 4 • October-December 2020

MRF Model-Based Estimation of Camera Parameters and Detection of Underwater Moving Objects

Susmita Panda, Image and Video Analysis Lab, Department of ECE, ITER, Siksha 'O' Anusandhan (Deemed), Bhubaneswar, India

Pradipta Kumar Nanda, Image and Video Analysis Lab, Department of ECE, ITER, Siksha 'O' Anusandhan (Deemed), Bhubaneswar, India

ABSTRACT

The detection of underwater objects in a video is a challenging problem particularly when both the camera and the objects are in motion. In this article, this problem has been conceived as an incomplete data problem and hence the problem is formulated in expectation maximization (EM) framework. In the E-step, the frame labels are the maximum a posterior (MAP) estimates, which are obtained using simulated annealing (SA) and the iterated conditional mode (ICM) algorithm. In the M-step, the camera model parameters, both intrinsic and extrinsic, are estimated. In case of parameter estimation, the features are extracted at coarse and fine scale. In order to continuously detect the object in different video frames, EM algorithm is repeated for each frame. The performance of the proposed scheme has been compared with other algorithms and the proposed algorithm is found to outperform.

KEYWORDS

EM algorithm, Feature Extraction, Iterated Conditional Mode, MAP Estimation, Model Camera Calibration, Multi-Resolution framework, Simulated annealing, Spatio Temporal Markov Random Field Model

1. INTRODUCTION

The problem of underwater video object detection has received considerable attention during last decades and appreciable progress has been made in this direction (Emberton, Chittka, & Cavallaro, 2018; Hossain, Alam, Ali, & Amin, 2016; Mohapatra, Mahapatra, Mahapatra, & Swain, 2015; Walther, Edgington, & Koch, 2004).The underwater moving object suffers from limited range of visibility, low contrast, non-uniform lighting, blurring, bright artefacts, colour diminished and noise (Ancuti, Ancuti, Haber, & Bekaert, 2012; Emberton et al., 2018; Zhang et al., 2017). An automated system for detection and tracking of underwater moving objects has been developed, which of interest to the oceanographic researchers ((Mohapatra et al., 2015). Variable lighting condition and the presence of noise from high contrast debris pose challenge for object detection and tracking. Walther et al. (Walther et al., 2004) have proposed a novel method to overcome the above issues. Negrea et al. (Negrea, Thompson, Juhnke, Fryer, & Loge, 2014) have presented an adaptive background subtraction algorithm for detection and motion prediction which is used for tracking. Design of this fully automated system removes the frames without any activity and hence there is cost reduction for fish monitoring.

This problem of underwater object detection can be of two types. In first case, the object moves while the camera is static, and in second case, both the object and camera are in motion. Often in real

DOI: 10.4018/IJCINI.2020100101

world scenario, the second case is more prevalent and challenging than the first one. This is valid in a real-world scenario, where neither the camera model parameters nor the object is known, this motivated us to address the issue in this research work.

In this paper, we have attempted to detect the underwater video objects under varying illumination condition. The problem is formulated as an incomplete data problem and the Expectation and Maximization (EM) approach has been adopted to solve the problem. Our main contributions are: (i) three new Spatio Temporal MRF models for classification of pixel labels in the E step, (ii) new features based model parameter estimation using pipelining approach in the M step, (iii) a continuous Underwater video object detection scheme using EM framework, and (iv) the EM algorithm in Multiresolution framework. In the proposed framework, no a priori knowledge of the camera model parameters is necessary. In E-step of the EM algorithm, the video object is segmented based on the video frame model. The problem of frame label estimation is formulated as a Maximum a posterior (MAP) estimation problem and these MAP estimates are obtained by an algorithm which is a combination of Simulated Annealing (SA) and the Iterated Conditional Model (ICM) algorithm. Subsequently, in M-step, the estimated frame labels are used to estimate the intrinsic and extrinsic parameters of the camera model. The proposed features are extracted from the labelled frames and weighted appropriately before being fed to the pipeline. These weighted corner features are used to estimate the camera intrinsic and extrinsic parameters using the 2D optimization method (Zhou, Cui, Peng, & Wang, 2012). E step and M step are repeated to continuously detect the video objects with the moving camera. The camera calibration errors have been computed and the estimated parameters are chosen based on the minimum calibration error. The segmentation accuracy has been validated by four quantitative measures. The advantage of the proposed multiresolution framework is that the execution time substantially reduced as compared to considering the fine scale images. The performance of the proposed algorithm has been compared with the Stolkin's E-MRF algorithm (Liu, Dai, Wang, Zheng, & Zheng, 2016; Prabowo, Hudayani, Purwiyanti, Sulistiyanti, & Setyawan, 2017; Rustam Stolkin, Greig, Hodgetts, & Gilby, 2008) algorithm and found to be superior.

Rest of the paper is organized as follows. The related research works are presented in Sec. 2. The proposed schemes are discussed in Sec. 3 and the EM framework is presented in Sec. 4. The new Spatio Temporal MRF models are presented in the Estep of Sec. 5. The M-step with the proposed weighted feature along with parameter estimation algorithm is provided in Sec. 6. Results and necessary discussions are presented in Sec. 7. Finally, conclusions are presented in Sec. 8.

2. RELATED WORK

Capturing the underwater object motion with the camera inside water is a real world problem and has been studied in detail by Amanda et al.(Silvatti et al., 2013) . Recently, the notion of multi view geometry, specifically two views have been employed for two new formulations, one for global optimal solution and the other for outliers (Kang, Wu, Wei, Lao, & Yang, 2017). Underwater object tracking in real time is often necessary and toward this end few research efforts have been directed for practical applications (Cho, Jung, Lee, Rim, & Lee, 2016; D. Zhang, Kopanas, Desai, Chai, & Piacentino, 2016) .

Many real-world underwater object detection problem have been addressed using EM framework (Chandan & Bala, 2009; Dempster, Laird, & Rubin, 1977). In the E-step of the EM framework, image modelling plays a crucial role. In this regard, Markov Random Field (MRF) model has been extensively used as the a priori model of the image labels (Geman & Geman, 1987; Li, 1994). Iterated Conditional Mode (ICM) algorithm is used (Besag, 1986) for simultaneous estimation of MRF model parameters and the image labels. Stolkin et al. (Stolkin, Hodgetts, Greig, & Gilby, 2007) have proposed Extended MRF (E-MRF) based model considering the interaction between the pixels of the observed frame and the corresponding pixels of predicted frame. The authors have used the E-MRF model develop a tracking algorithm that simultaneously estimates camera model

parameters and the class labels of video sequences (Rustam Stolkin et al., 2008). Recently, in order to improve the performance of Stolkin's algorithm, Panda et al. (Panda & Nanda, 2015) have proposed weighted oriented feature-based camera model parameter estimation and object detection. In order to detect moving object underwater H. Liu et. al (Liu et al., 2016) propose an approach which combine background subtraction and three frame difference considering the camera to be fixed. Similarly (Prabowo et al., 2017) addressed an adaptive background modelling method to detect moving objects on an underwater video. It has been observed that time varying background in a video sequence poses a challenge which has been addressed by Kalyan et al. (Halder, Tahtali, & Anavatti, 2016). They have identified and tracked the moving objects by the thresholding algorithm and Regression Neural Network. A new fish detection algorithm has also been implemented to identify and localize fish occurrences in each frame, under partial occlusion, and amidst dynamic texture patterns formed by whirls of sand on the seabed (Boudhane & Nsiri, 2016).

In the M-step of EM framework, the accuracy of estimation of camera parameters greatly depends on proper selection of features. The improved Harris corner detection algorithm has been used to extract features (Qiao, Tang, & Li, 2013) with reduced time for detection. As far as parameter estimation is concerned, Zhang (Z. Zhang, 2000) has proposed a closed form solution-based technique where the camera parameters can be estimated using the observed planar pattern which may move freely. Usually, camera calibration involves two steps; the first step is the linear computation of initial parameters values followed by the computation of the final parameters by nonlinear optimization. As an extension, Heikkila et al. (Heikkila & Silven, 1997) have proposed a four step procedure which is an extension of the two step method. Zhou et al. (Zhou et al., 2012) have proposed a novel optimization algorithm for estimating camera parameters by minimizing the distance error between calculated point and the real point in 3D measurement coordinate system. They have employed Levenberg-Marquardt algorithm to update the camera parameters. By and large, EM algorithm has been employed to simultaneously estimate camera parameters and pixel labels of frames to continuously detect the moving object with the camera in motion. Our proposed scheme presented in Sec.3 is based on EM framework.

3. PROPOSED SCHEME

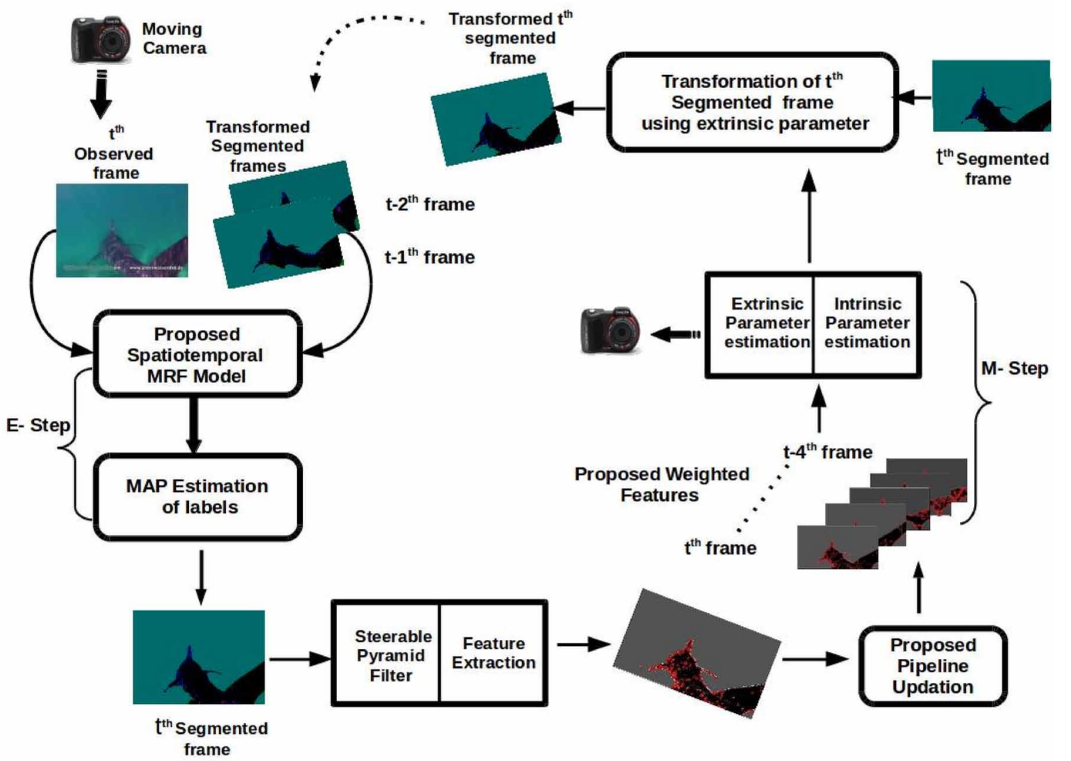
We have considered the underwater object detection problem when both camera and the object are in motion. Since both camera and object are in motion, the estimation of camera model parameters needs the previously available segmented frames. The segmentation of a given frame depends upon the a priori estimated camera parameters. Thus, it is conceivable that estimation of the image labels and the camera parameters are interdependent. In this regard, we have cast the problem as an incomplete data problem and employed the Expectation and Maximization (EM) algorithm which is presented in Figure 1.

Initially, the current observed frame together with the previously available segmented frames are used for spatio temporal MRF modelling. The segmentation problem is formulated as pixel labelling problem and the pixel labels are estimated using MAP estimation criteria. The MAP estimates yield the current segmented frame and the features that have been extracted from this frame are used with the features of the previous frames available in the pipeline. Features corresponding to different frames are used to estimate the camera model parameters. The extrinsic parameters of the camera model are used to transform the current segmented frame and the transformed frames are used for spatio temporal modelling of the subsequent frame. This process of estimation of frame labels and camera parameters is repeated for object detection till all the frames are exhausted.

The proposed scheme in EM framework is presented in Figure 2.

In the E-step, the frame labels have been estimated using MAP estimation principle while in the M-step the camera model parameters are estimated. This is shown in Figure 2, where the given frame is modelled as spatio temporal MRF and the MAP estimation is obtained using the simulated annealing and ICM algorithm. The camera model parameters, both extrinsic and intrinsic ones, are

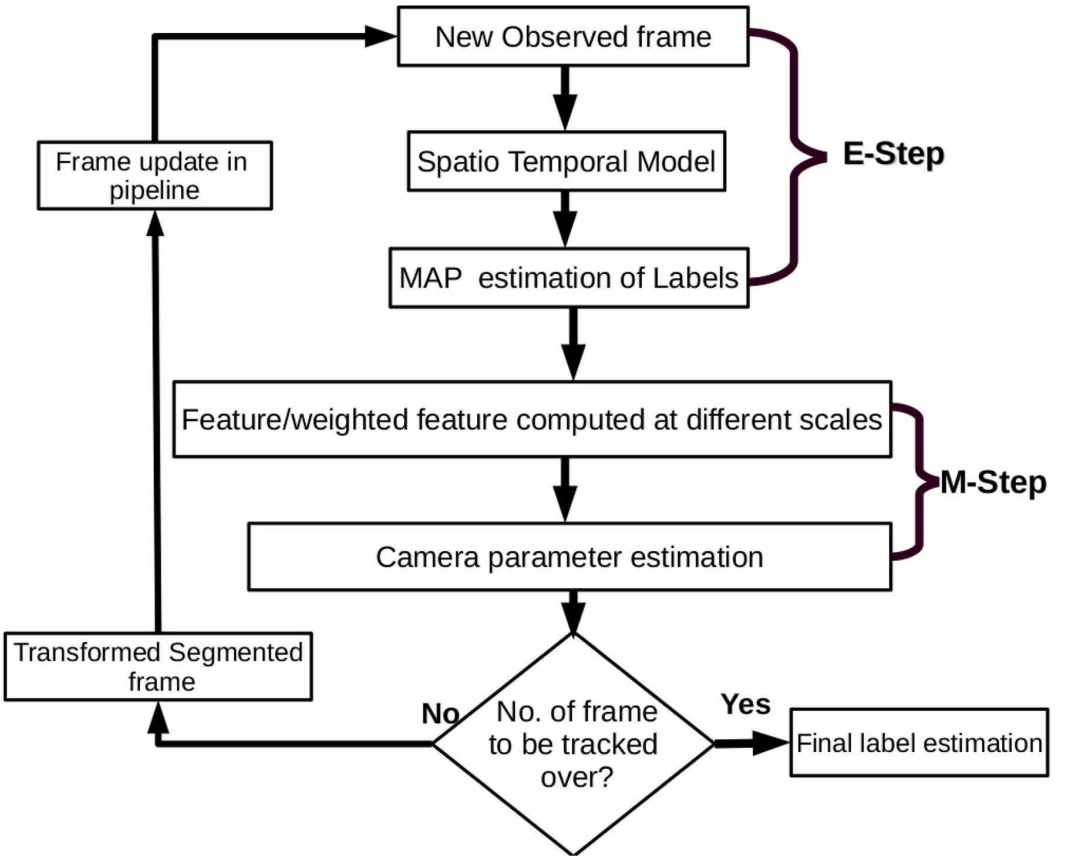
Figure 1. Schematic representation of the proposed scheme using EM framework



estimated using the features/weighted features computed using the segmented frames. The parameter estimation step works using the notion of pipelining. The parameters are estimated using the features of the current frame as well as the features of previously available transformed frames. These transformed frames are obtained by convolving the prior segmented frames with the corresponding estimated rotational parameters. The features of the transformed frames and the current frame have been used to estimate the intrinsic and extrinsic parameters. Using these estimated parameters, the segmentation is obtained by the MAP estimation. In parameter estimation step, we need features of five frames to estimate the camera model parameters. As shown in Figure 3,

at time instant 't' all the slots have been filled with the features of different frames and the parameters have been estimated based on the available features. Because of availability of the features of all frames, the error in the parameter estimation is expected to be low. At $\left(t+1\right)^{th}$ time the next features of frames are available and accordingly the parameters will be estimated accurately, particularly the extrinsic parameters will be updated while the intrinsic parameters will not change. The parameters estimated with the available features in the pipeline are obviously not the correct estimates and the estimates are presented in Table 1 and Table 2. Subsequently, in the next time slot, the features of the view i.e view 1 are shifted to the next stage and the features of view 2 are input to the first stage of the pipeline. The camera parameters thus estimated are tabulated in Table 1 and Table 2.

Figure 2. Block diagram of the proposed approach



4. EM FRAMEWORK

The EM algorithm, as applied for the incomplete data problem consists of two steps, the Expectation step and the Maximization step. In the Expectation step, estimate of the complete data is obtained from the incomplete data, while in the Maximization step, these estimated complete data are used to maximize the likelihood estimate of the camera model parameters. These two steps are alternated till the convergence. At convergence, the camera parameter estimates are the maximum likelihood estimates and the label estimates are the MAP estimates. In the E step, the expectation of the joint probability distribution of the observed image X and the unobserved label Z given the observed image X and the current estimates of the model parameter θ.

That is in E step, the following is evaluated:

$$E\left[\log_e P\left(X,Z \mid \theta\right) \mid X,\hat{\theta}^n\right] \tag{1}$$

In the M-step, the parameter vector θ is estimated to be θ^{n+1} by maximizing the expectation of this joint probability. This is tantamount to maximizing the likelihood function $P\left(X,Z|\theta^{n+1}\right)$ given the observed image X and θ^n. Since log is a monotonic function, often in practice the likelihood

Figure 3. Notion of pipelining for camera model parameter estimation

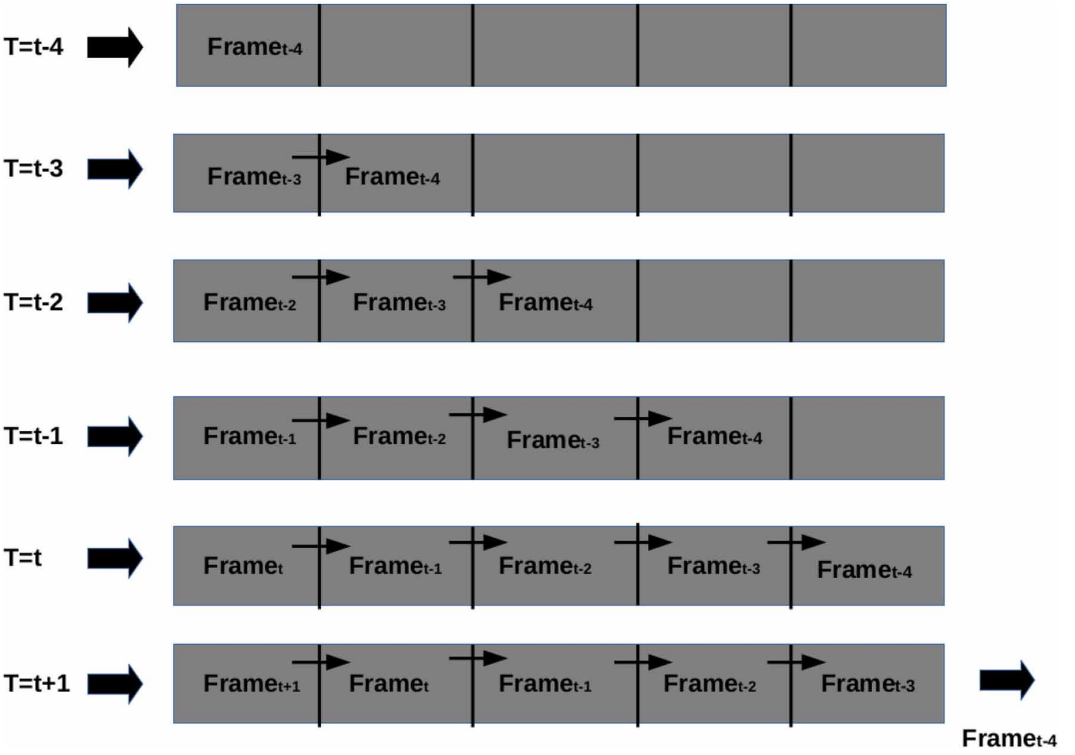


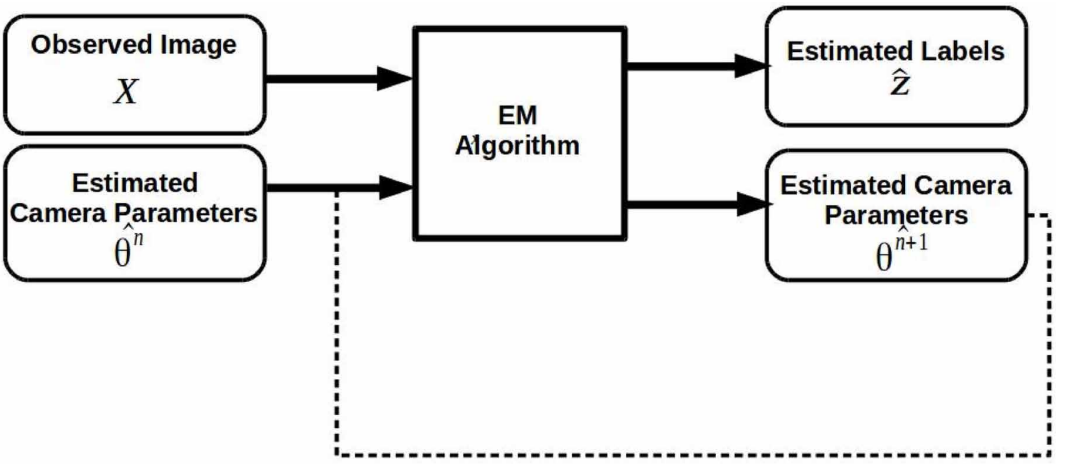
Table 1. Intrinsic Parameter estimation with the notion of pipelining. (Frame 15-22 of Whalesharks in Philippines southern Leyte, Underwater video)

Frame Time	View	f_x (mm)	f_y (mm)	u_0 (mm)	v_0 (mm)
F_{t-4}	View 0	*	*	*	*
	View 1	*	*	*	*
	View 2	*	*	*	*
	View 3	*	*	*	*
	View 4	*	*	*	*
F_{t-1}	View 0	192.67	174.37	-23.11	32.039
	View 1	192.67	174.37	-23.11	32.039
	View 2	192.67	174.37	-23.11	32.039
	View 3	192.67	174.37	-23.11	32.039
	View 4	*	*	*	*
F_t	View 0	36.31	38.15	189.17	40.62
	View 1	36.31	38.15	189.17	40.62
	View 2	36.31	38.15	189.17	40.62
	View 3	36.31	38.15	189.17	40.62
	View 4	36.31	38.15	189.17	40.62

Table 2. Extrinsic Parameter estimation with the notion of pipelining. (Frame 15-22 of Whalesharks in Philippines southern Leyte, Underwater video)

Frame Time	View	$\theta(deg.)$	t_x (mm)	t_y (mm)	t_z (mm)	Error of calibration
F_{t-4}	View 0 View 1 View 2 View 3 View 4	* * * *	* * * * *	* * * * *	* * * * *	* * * * *
F_{t-1}	View 0 View 1 View 2 View 3 View 4	93.25 11.54 13.09 23.95 *	308.50 127.84 261.16 304.42 *	177.87 .275 -.62 38.57 *	775.82 294.89 558.32 678.93 *	50.93 15.37 39.92 45.9 0.00
F_t	View 0 View 1 View 2 View 3 View 4	18.9 10.24 9.42 4.2 6.33	-448.88 -565.16 -291.84 -381.21 -412.91	30.44 124.46 -18.17 -26.08 4.12	116.95 151.27 82.30 113.31 119.09	34.86 48.58 65.65 28.74 35.76

Figure 4. Recursive estimation of camera parameter



function is maximized instead of actual function. Thus, in M step, θ^{n+1} is obtained by maximizing the following:

$$E\left[log_e P(X,Z|\hat{}^{n+1} \mid X,\theta^n\right] \tag{2}$$

This estimated $\hat{\theta}^{n+1}$ is used in the E step to estimate the labels $\hat{z}$ at $\left(n+1\right)^{th}$ instant and the process is repeated till convergence as shown in Figure 4.

5. E-STEP

The hidden value i.e. the labels of the image "z" is estimated in the Expectation step. This is obtained by determining the following expected value.

$$E\left[log_e P\left(X, Z \mid \theta\right) \mid X, \theta^k\right] \tag{3}$$

The $log_e P\left(X, Z|\theta\right)$ is evaluated as follows. $P\left(X, Z|\theta\right)$ is evaluated for every individual pixel and then summed over the entire image. For a given pixel at $\left(i; j\right)^{th}$ location, the above joint density is evaluated as

$$P\left(X_{i,j}, Z_{i,j}|\theta^k\right) = P\left(X_{i,j}|Z_{i,j}, \theta^k P\left(Z_{i,j}|\theta^k\right)\right) \tag{4}$$

In Equation (4) $P\left(Z_{i,j}|\theta^k\right)$ is the prior probability distribution function and assuming Z to be Spatio Temporal MRF, the prior probability $P\left(Z_{i,j}|\theta^k\right)$ can be expressed as Gibbs distribution i.e

$$P(z_{i,j}|\theta^k) = \frac{e^{-U_{z_{i,j}}/T}}{Z'} \tag{5}$$

We assume the observed image process X is obtained from Z by a Gaussian degradation process. Hence $P\left(X = x \mid Z = z_{ij}, \theta^k\right)$ can be expressed as

$P\left(X_{ij} = z_{ij} + n_{ij} \mid Z = z_{ij}, \theta^k\right)$ or $P\left(n_{ij} = x_{ij} - z_{ij} \mid Z = z_{ij}, \theta^k\right)$

Assuming n_{ij} to be Gaussian,

$$P\left(X_{i,j} \vee Z_{i,j}, \theta\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left\{\frac{-\left(x_{ij} - z_{ij}\right)^2}{2\sigma^2}\right\}} \tag{6}$$

Therefore,

$$P\left(X_{i,j}, Z_{i,j}|\theta^k\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left\{\frac{-\left(x_{ij} - z_{ij}\right)^2}{2\sigma^2}\right\}} - \frac{e^{-U_{z_{i,j}}/T}}{Z'} \tag{7}$$

Considering $Z^{`}$ as constant and T as unity, and taking logarithm of both sides of Equation (7),

$$\sum_{i,j\epsilon I} log_e P\left(X_{i,j}, Z_{i,j} | \theta^k\right) =$$

$$\sum_{i,j\epsilon k} \gamma_1 \left[f\left(Z_{i,j}, Z_{i+m,j+n}\right)\right] + \gamma_2 \left[f\left(Z_{i,j}, \hat{Z}_{i,j}\right)\right] \frac{1}{2} log\left(\sigma^2\right) + \frac{\left(X_{i,j} - z_{i,j}\right)^2}{2\sigma^2} \tag{8}$$

Where, γ_1 and γ_2 are the weighting parameters fixed to deal with the poor visibility condition and f (.) denotes the pairwise clique potential function. Equation (8) is minimized to obtain $\hat{z}$. The value of $\hat{z}$ has been obtained by SA and ICM algorithm. With this estimated $\hat{z}$ labels the complete likelihood function $E\left[log_e P\left(X, Z|\theta\right)|X, \theta^k\right]$ has been maximized to obtain the estimate of parameter vector θ.

We have proposed the following three Spatiotemporal MRF models as the prior model for the image label Z.

5.1. Proposed Spatiotemporal MRF (ST-MRF) With First Order Spatial and Temporal Neighbourhood (1st Model)

In the E-step of the algorithm, the segmentation problem has been cast as a pixel labelling problem and the label estimates are obtained in MAP framework. In this context, the evaluation of the posterior probability requires the knowledge of the a priori model of the labels and the degradation process of the labels. The degradation process is assumed to be Gaussian and the a priori pixel label model is the spatiotemporal MRF model. The temporal frames are the previously transformed frames. Thus, the current frame at t and the transformed frames corresponding to (t-1) and (t-2) time instants are used for 2nd order spatiotemporal modelling. Analogously, first order spatiotemporal modelling uses the t^{th} frame together with the transformed frame at $\left(t-1\right)^{th}$ instant. In the following, we present the spatiotemporal MRF modeling with first order Spatial and first order temporal neighbourhood.

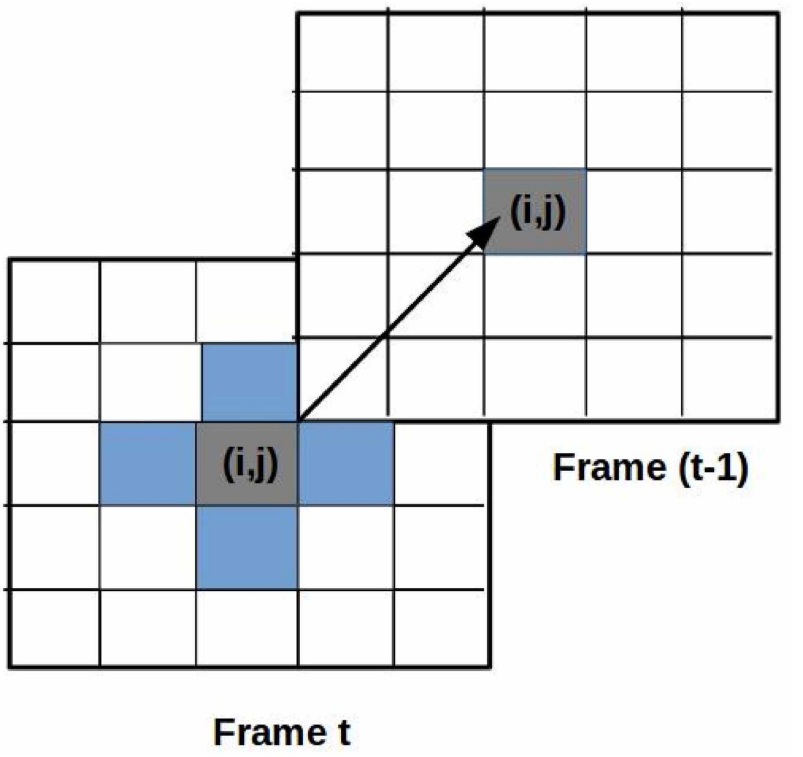
Since observed video sequence x is a 3D volume consisting of image frames in temporal direction, x_t denotes the frame at time t, and x_{st} denotes a site s of the temporal frame x_t. Therefore, x_{st} corresponds to the spatio- temporal coordinate of the grid (s, t). Let z denotes the segmentation of video sequence x and z_t is the segmentation of the x_t^{th} frame. Z_t has been assumed to be MRF and z_t is a realization of Z_t. This assumption of Markovianity is in the spatial direction. We have also assumed Markovianity in the temporal direction. As shown in Figure 5, we have shown one temporal frame at (t -1) time instant for considering 1storder temporal neighbourhood. Since we have assumed to have the Markov Model in both spatial and temporal direction, the Markovianity is satisfied for both spatial and temporal directions as well. In spatial direction

$$P\left(Z_{st} = z_{st} \mid Z_{rt} = z_{rt}; \forall r\epsilon \left(S\right), s \neq r\right)$$

$$P\left(Z_{st} = z_{st} \mid Z_{rt} = z_{rt}; \left(r\right)\epsilon \eta_{st}\right) \tag{9}$$

Where, η_{st} denotes the neighborhood of $\left(s,t\right)$ and $S = \left(M \times N\right)$ denotes the lattice of z_t. Figure 5 shows the schematic representation of Spatio Temporal modeling but the local property in temporal direction is given as

Figure 5. Spatio Temporal MRF model with first order spatial and temporal neighbourhood



$$P\left(Z_{st} = z_{st} \mid Z_{sq} = z_{sq}; q \neq t; \forall \left(s,q\right) \epsilon V\right)$$

$$P\left(Z_{st} = z_{st} \mid Z_{sq} = z_{sq}, \left(s,q\right) \epsilon \eta_{st}\right) \tag{10}$$

Where, V denotes the 3D volume of the video sequence. The priori probability can be expressed as Gibbs distribution and can be expressed as

$$P\left(Z_t = z_t\right) = \frac{e^{-U\left(z_t\right)/T}}{Z'} \tag{11}$$

Where, $U\left(z_t\right)$ is the energy function which can be expressed as

$$U\left(z_t\right) = \sum_{\alpha \epsilon C} V_{sc}\left(z_t\right) + \sum_{\alpha \epsilon C} V_{tc}\left(z_t\right) \tag{12}$$

Where, $V_{sc}\left(z_t\right)$ denotes the clique potential in spatial domain while considering a single frame and is given by

$$V_{sc}\left(z_t\right)=\begin{cases}+\alpha & if z_{st}\neq z_{rt},\left(r,t\right)\epsilon S\\ 0 & if z_{st}=z_{rt},\left(r,t\right)\epsilon S\end{cases} \tag{13}$$

Similarly, the clique potential function in temporal direction with first order neighbourhood can be expressed as

$$V_{tc}\left(z_t\right)=\begin{cases}+\beta & if z_{st}\neq z_{s,t-1},\left(s,t-1\right)\epsilon V\\ 0 & if z_{st}=z_{s,t-1},\left(s,t-1\right)\epsilon V\end{cases} \tag{14}$$

Where, $V_{tc}\left(Z_t\right)$ denotes the clique potential function in temporal direction.

5.2. Proposed Spatio Temporal MRF (ST-MRF) with First Order Spatial Neighbourhood (SN) and 2nd Order Temporal Neighbourhood (TN) (2nd Model)

Figure 6 shows the spatiotemporal MRF with first order Spatial neighbourhood and 2nd order temporal neighbourhood.

The first order spatial neighbourhood is presented in Equation 9 and for Figure 6 the local characteristic corresponding to temporal direction Markovianity is given by

$$P\left(Z_{st}=z_{st}\mid Z_{sq}=z_{sq};q\neq t;\forall\left(s,q\right)\epsilon V\right)$$

$$P\left(Z_{st}=z_{st}\mid Z_{sq}=z_{sq},\left(s,q\right)\epsilon\eta_{st}\right) \tag{15}$$

Where, η_{st} refers to the 2nd order neighbourhood structure. The energy function is given by

$$U\left(z_t\right)=\sum_{\alpha\epsilon C}V_{sc}\left(z_t\right)+\sum_{\alpha\epsilon C}V_{tc}\left(z_t\right) \tag{16}$$

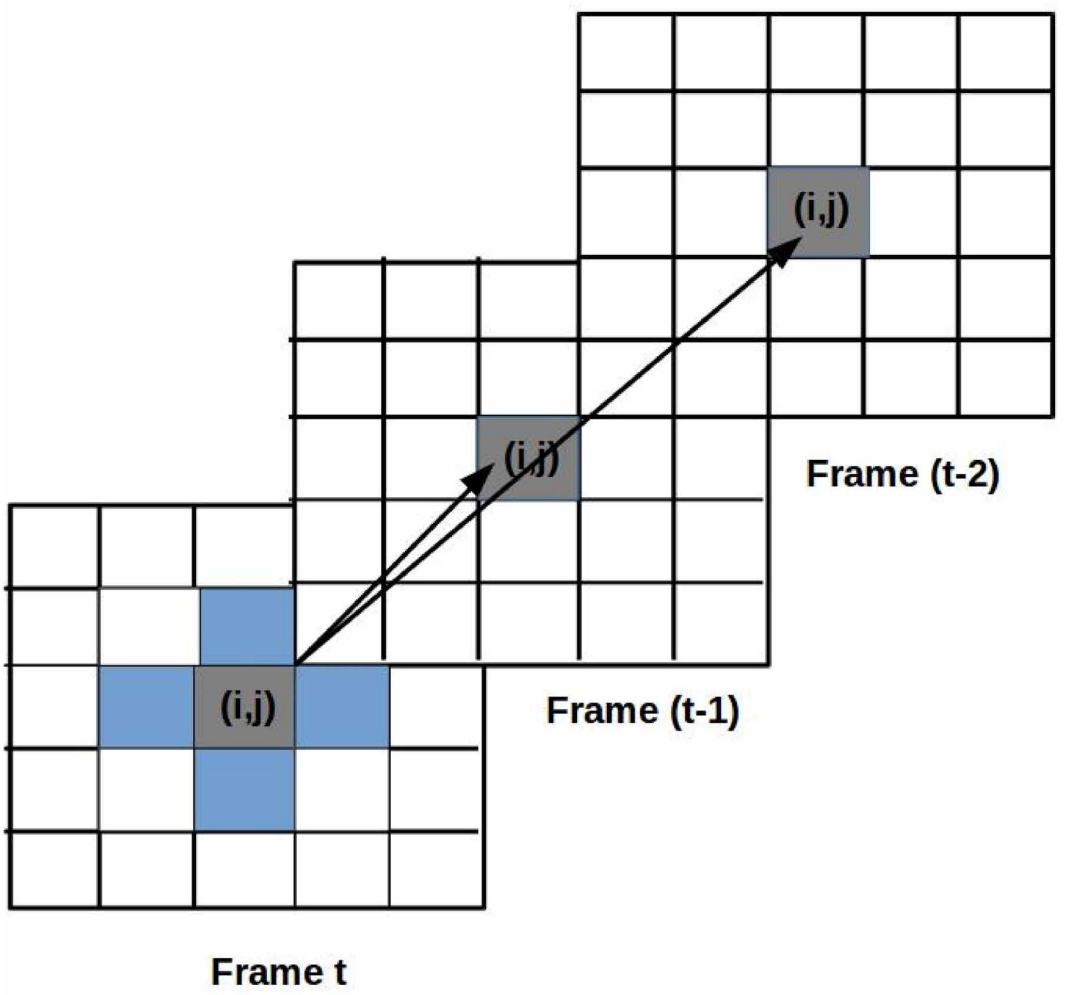
$V_{sc}\left(z_t\right)$ is same as defined in Equation 13 whereas there is clique potential functions which takes care of first order and 2^{nd} order terms. Hence, $V_{tc}=V_{tc1}+V_{tc2}$. V_{tc1} corresponds to first order term and is given by

$$V_{tc1}\left(z_t\right)=\begin{cases}+\beta & if\ z_{st}\neq z_{s,t-1},\left(s,t-1\right)\epsilon V\\ 0 & if\ z_{st}=z_{s,t-1},\left(s,t-1\right)\epsilon V\end{cases} \tag{17}$$

and for 2^{nd} order term the potential function is defined as

$$V_{tc2}\left(z_t\right)=\begin{cases}+\beta & if\ z_{st}\neq z_{s,t-2},\left(s,t-2\right)\epsilon V\\ 0 & if\ z_{st}=z_{s,t-2},\left(s,t-2\right)\epsilon V\end{cases} \tag{18}$$

Figure 6. Spatio Temporal MRF with 1st order spatial neighbourhood and 2nd order temporal neighbourhood



5.3. Proposed Spatio Temporal MRF (ST-MRF) With 2nd Order Spatial Neighbourhood (SN) and 2nd Order Spatio Temporal Neighbourhood (STN) (3rd Model)

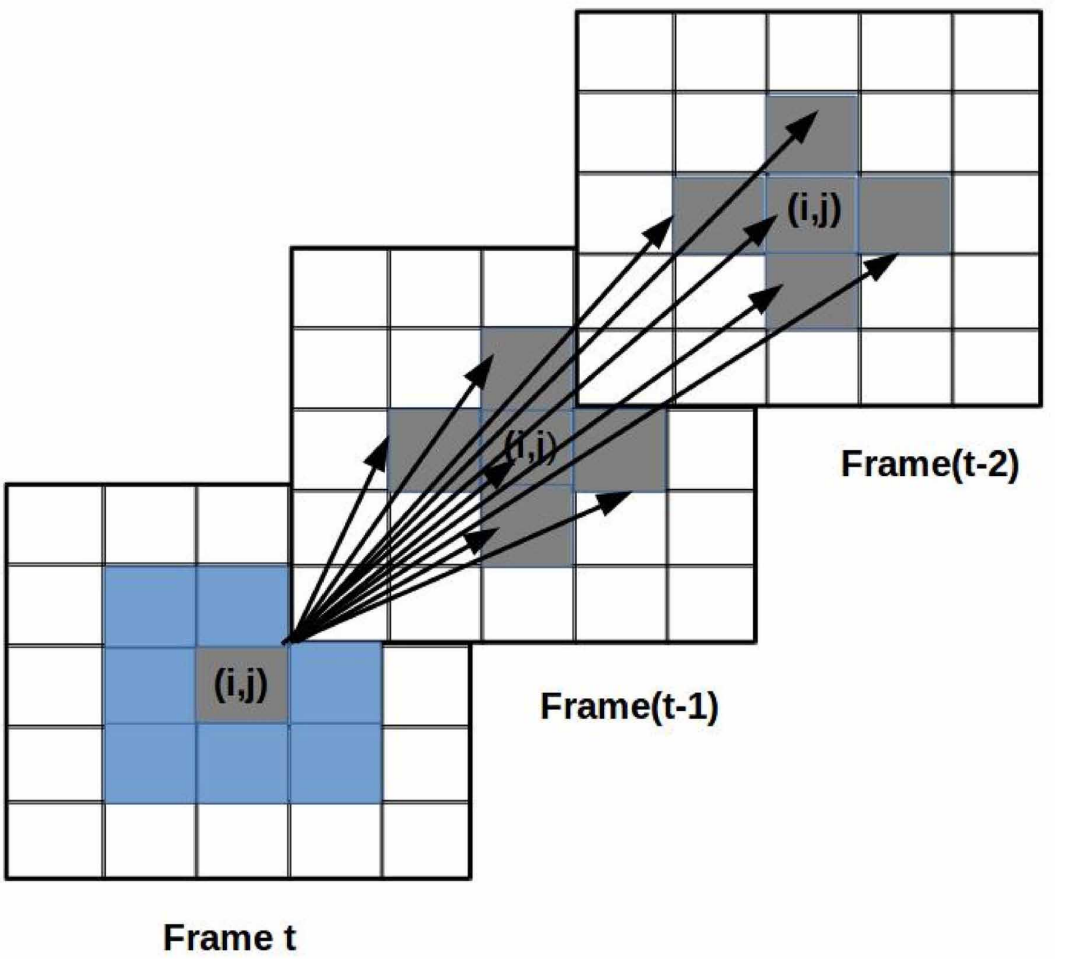
For Spatial model, the local characteristic is same as Equation 9 but the local characteristic for the 2nd order neighbourhood structure is given by:

$$P\left(Z_{st} = z_{st} \mid Z_{rq} = z_{rq}; r \neq s, q \neq t; \forall \left(r,q\right) \epsilon V\right)$$

$$P\left(Z_{st} = z_{st} \mid Z_{rq} = z_{rq}, \left(r,q\right) \epsilon \eta_{st}\right) \tag{19}$$

Where, η_{st} denotes the 2^{nd} order neighbourhood structure. Figure 7 shows the Spatio Temporal structure with second order neighbourhood. The energy function $U\left(z_t\right)$ is given by

Figure 7. Spatio Temporal MRF with 2nd order spatial neighbourhood and 2nd order Spatio Temporal Neighbourhood



$$U\left(z_t\right)=\sum_{\alpha\epsilon C}V_{sc}\left(z_t\right)+\sum_{\alpha\epsilon C}V_{tc}\left(z_t\right) \tag{20}$$

The clique potential function for the first order neighbourhood will be same as Equation 12, but for second order neighbourhood the potential function consists of two interactions V_{tc1} and V_{tc2} respectively for first order and 2^{nd} order neighborhood structure.

For the sake of clarity, they are separately presented as follows:

$$V_{tc1}=\begin{cases}+\beta & if\ z_{st}\neq z_{r,t-1},\left(s,t\right),\left(r,t-1\right)\epsilon V\\ 0 & if\ z_{st}=z_{r,t-1},\left(s,t\right),\left(r,t-1\right)\epsilon V\end{cases} \tag{21}$$

$$V_{tc2} = \begin{cases} +\beta_1 & if\ z_{st} \neq z_{r,t-2}, \left(s,t\right), \left(r,t-2\right) \epsilon V \\ 0 & if\ z_{st} = z_{r,t-2}, \left(s,t\right), \left(r,t-2\right) \epsilon V \end{cases} \tag{22}$$

6. M-STEP

We estimate the camera model parameters in the M-step.

$$\hat{\theta}^{k+1} = argmax_{\theta} E[\log_e P(X, \hat{Z} \mid \theta) \mid X, \theta^k] \tag{23}$$

The image labels $\hat{z}$ have been estimated in E step and these estimated labels have been used together with the observed frame X_t at t^{th} time instant to determine the estimate of the camera model parameters. The 2D optimization based method as proposed by Zhou et al. (Zhou et al., 2012) has been used to estimate the camera model parameters. This is tantamount to optimizing the likelihood function of 23. The observed frame X_t and the estimated labels $\hat{Z}_t$ are used in the 2D optimization method.

The feature points considered are the Harris corner points of the whale, which is shown in Figure 8. These corner points are mapped into the camera coordinate plane and in the sequel to the image coordinate plane. We have not considered distortion and therefore the distance between the estimated image point in the image coordinate plane $\hat{i}_u$ and real image point i_u is minimized. Hence the estimated point in the image coordinate $\hat{i}_u$ is a function of intrinsic parameters $\left(f_x, f_y, u_0, v_0\right)$ and extrinsic parameters *R* and *t* i.e. $\hat{i}_u = f\left(f_x, f_y, u_0, v_0, R, t\right)$ and the parameter vector $\theta = \left(f_x, f_y, u_0, v_0, R, t\right)^T$. Therefore, in this case the problem is reformulated as

$$\hat{\theta} = \underset{\theta}{\arg\min} \sum_{i=1}^{M} \sum_{j=1}^{N} i_u - \hat{i}_u^{\,2} \tag{24}$$

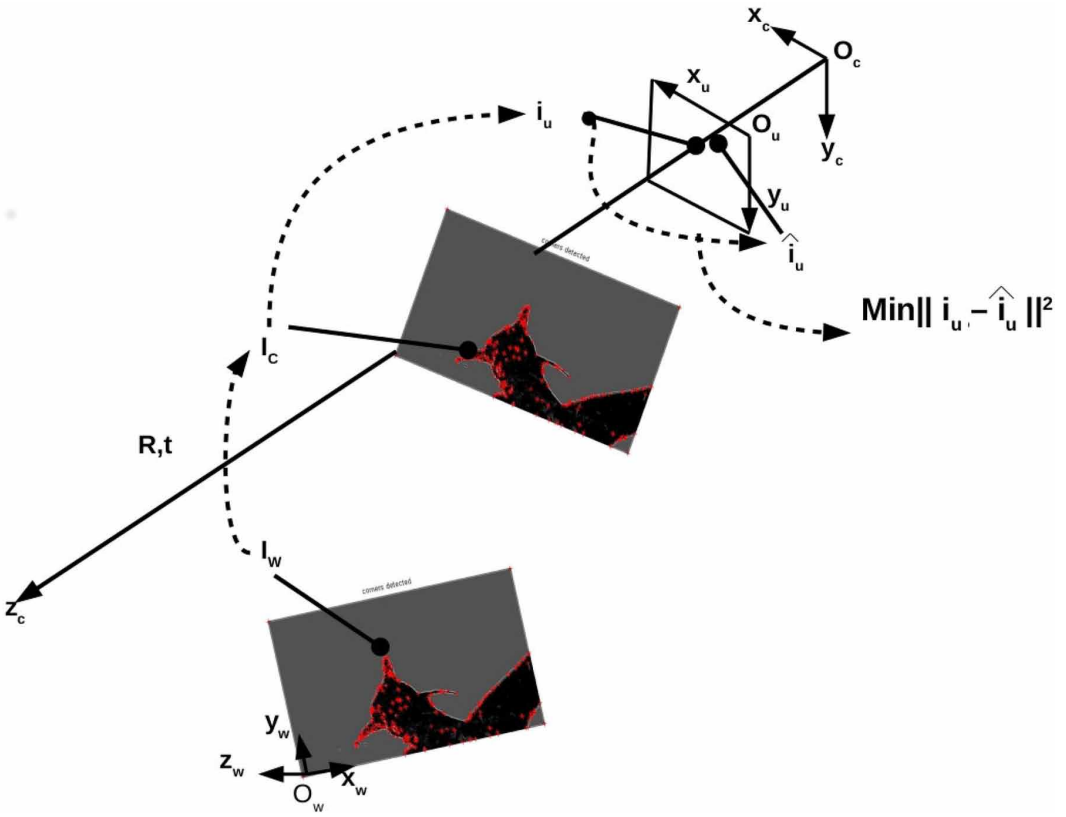
The estimated $\hat{\theta}$ provides us the estimated parameters $\left(\hat{f}_x, \hat{f}_y, \hat{u}_0, \hat{v}_0, \hat{R}, \hat{t}\right)$. These parameters are used to transform the a priori segmented image to be used with the current image for spatiotemporal MRF modelling which is used to obtain the estimation of the labels in the E-step.

6.1. Proposed Weighted Feature for Parameter Estimation

The accuracy of the estimation of camera parameters greatly depends upon the appropriate feature points, which is evident from previous section. The movement of the object is in underwater and hence, the movement in each frame may lead to improper extraction of feature points. In order to extract proper feature points, we have used steerable pyramid filters with different angles for different frames. Steerable filters have been employed to obtain different features of a given frame with different orientation. This filter is recursive in nature and hence the k directional bandpass filter can be expressed as

$$B_m\left(u,v\right) = HP\left(f_1, f_{\frac{N}{2}}, s\right) cos^{k-1}\left(\theta - \frac{m\pi}{k}\right) \tag{25}$$

Figure 8. This figure shows the step for 2D optimization of whale. Where I_w = World coordinate plane, I_c = Camera coordinate plane, i_u = image plane, O_c = optical center of the camera and z_c = optical axis of the camera lens.



Where m = 0…k-1

$$S = \left(u^2 + v^2\right)^{\frac{1}{2}} \tag{26}$$

S is the radial variable in frequency space and $\theta = tan^{-1}\left(\frac{v}{u}\right)$ is the angular variable in frequency space.

HP(a,b,f) is a high pass transfer function, raised to cosine.

$$\begin{cases} 0; f \le a \\ \sqrt{\frac{1}{2}\left[1 - cos\left[\pi\left(\frac{f-a}{f-b}\right)\right]\right]}; a < f < b \\ 1; f \ge b \end{cases} \tag{27}$$

The Kernels at different angle have been applied to the considered frames for feature extraction.

Harris corner detection algorithm has been the choice for detection of corner points. From a practical standpoint, it may be conceivable that accurate corners may correspond to sub-pixel coordinate positions instead of pixel coordinates. Hence, we have adhered to the improved Harris corner sub-pixel corner detection algorithm (Qiao et al., 2013) in different video frames. For a given frame, the corner points are weighted to take care of the orientation and movement. We have assigned different weight age to different frame's feature points with a view to take care of the movements in different frames. We have extracted features at coarse resolution which are used for parameter estimation. The proposed framework with different models are used to obtain segmentation. Gaussian Pyramid (Karasaridis & Simoncelli, 1996) has been constructed based on following.

$$P_{Gaussian}\left(I\right)_{n+1} = S\downarrow\left(G_{\sigma} **P_{Gaussian}\left(I\right)_{n}\right) \tag{28}$$

The operator $S\downarrow$ down-samples an image; the j, k^{th} element of $S\downarrow$ (I) is the 2 j, $2k^{th}$ element of I. The n^{th} level of a pyramid P(I) is denoted as $P\left(I\right)_{n}$. G_{σ} is a linear operator that takes an image to the convolution of that image with a Gaussian. The frame no. 19 at different resolution is shown in Figure 9.

7. RESULTS AND DISCUSSIONS

We have considered different viewsrom two data sets, the first one is Whalesharks in Philippines southern Leyte, Underwater video while, the second one is from Creepy chimera/Nautilus live video. Since, our pipeline consists of five stages, we have considered eight views (frames) from the first data set and five views from the second data set.

Features, particularly the corner features are extracted to be used for parameter estimation. The transformed segmented frames have been passed through Steerable pyramid filter to have maximum exposure of the edges of the object. Thereafter, the corner features have been extracted using the improved Harris corner detection algorithm. The features corresponding to frames 15, 16, 17, 18, and 19 are weighted with 0.4, 0.2, 0.1, 0.1, and 0.1, respectively. These weighted features in the pipeline are used to estimate the camera intrinsic and extrinsic parameters. The estimated intrinsic parameters corresponding to three image frame models are presented in Table 3. The corresponding calibration errors are tabulated in Table 4. As observed from Table 4 the calibration errors are less with oriented weighted features at coarse scale than those at fine scale.

Although this could be due to the reduced features at coarse scale, the minimum Calibration error is expected for accuracy of estimation of both intrinsic and extrinsic parameters. The intrinsic

Figure 9. (a) Finer image (480 × 270) (b) Image down sample to (240 × 135) (c) Image down sample to (120 × 67) (d) Image down sample to (60 × 33)

Table 3. Camera intrinsic parameters (in mm) with oriented weighted features using 2Doptimization at different scale for 1st, 2nd and 3rd Model

		1st Model				2nd Model				3rd Model			
Methods	**Scale**	f_x	f_y	u_0	v_0	f_x	f_y	u_0	v_0	f_x	f_y	u_0	v_0
Whale sharks in Philippines southern Leyte, Underwater video													
Stolkin's Model	**Fine**	36.31	38.15	189.17	40.62								
Oriented weighted features	**Fine**	36.39	37.62	174.60	38.72	37.12	35.82	156.43	38.72	37.85	35.83	156.43	39.72
	Coarse	37.48	38.29	188.59	42.44	37.48	38.29	188.59	39.65	37.48	39.14	189.89	38.27
Creepy chimera/Nautilus live video													
Stolkin's Model	**Fine**	32.12	37.71	69.48	291.1								
Oriented weighted features	**Fine**	32.32	32.11	75.60	297.1	32.78	33.71	72.91	297.1	33.52	33.95	73.63	297.9
	Coarse	35.22	37.54	59.57	269.8	35.12	38.08	60.34	270.7	33.28	39.64	60.89	270.1

Table 4. Camera calibration error (in pixels) using with oriented weighted features in different scale using 2D optimization for 1st, 2nd and 3rd model

	1st Model		2nd Model		3rd Model	
Original views	**Fine**	**Coarse**	**Fine**	**Coarse**	**Fine**	**Coarse**
Whale sharks in Philippines southern Leyte, Underwater video						
View1	5.67	0.82	5.20	0.71	4.98	0.65
View2	7.01	3.44	6.72	2.99	6.41	2.75
View3	5.88	4.17	5.27	3.72	5.04	3.68
View4	6.88	3.46	6.48	2.33	6.11	2.07
View5	5.39	2.35	5.26	2.04	4.99	1.88
Creepy chimera/Nautilus live video						
View4	4.02	3.40	3.33	3.01	3.02	2.01
View5	3.07	2.98	2.72	2.52	2.32	1.62

Table 5. Camera calibration error (in pixels) using 2D optimization for Stolkin's Model

Original views	**View1**	**View2**	**View3**	**View4**	**View5**
Whale sharks in Philippines southern Leyte, Underwater video					
	34.86	48.57	65.65	28.74	35.7
Creepy chimera/Nautilus live video					
	39.92	41.01	41.09	33.74	20.06

parameters for weighted oriented features are presented in Table 3. The estimated parameters are within the available focal length range i.e 4.5mm- 54mm. The calibration error is minimum for the third model at coarse scale level thus confirming the use of coarse scale. Comparing results of Table 4 and Table 5, the errors with the proposed models are very less than that of the Stolkin's model. As observed from Table 11, the execution time for the frames at coarse scale is in the range of (8-11) seconds.

The segmentation results obtained by the proposed algorithms and the existing three algorithms are presented in Figure 10 - Figure 19. Figure 10 - Figure 17 correspond to results of 8frames of video of whale sharks in Philippines southern Leyte, while, Figure 18 and Figure 19 correspond 32 and 33 frames of Creepy chimera/live video. Figure 10(f), 10(h) and 10(j) results correspond to finer scale while Figure 10(g), 10(i) and 10(k) correspond to coarse scale results. As observed from these figures for finer scale, there is an improvement with the 3rd model. As observed from Figure 10(f), there are misclassified pixels on the tail of the whale, and the number of misclassified pixels reduces with 2nd order neighbourhood structure based (ST-MRF) which is observed from Figure 10(j). This could be attributed to the model with weighted features. The reconstructed fine scale images from the coarse scales results are shown in Figure 10(g), 10(i) and 10(k). As observed from these figures, some portions of the boundary are blurred and the misclassified pixels are more in case of 1st order as compared to that of 2nd order ST-MRF case. This has been reflected on the Percentage of Misclassification Errors (PME) presented in Table 6.

The PME for Stolkin et al., Liu et al., and M. R. Prabowo et al. model is higher than those of the proposed models. Similar observations are also made for the results obtained for other frames presented in Figure 10 to Figure 17. This effect can also be observed from Table 6.

The proposed algorithms have also been tested for the second set of frames. The results corresponding to two frames have been presented in Figure 18 and Figure 19. As observed from the segmented results of Figure 18(d) - 18(i), sharp boundaries of the object could be retained in case of fine scale while the boundaries are blurred in case of coarse scales. Similar observations are also made for the results presented in Figure 19. Further, it is also observed that the set of results obtained

Figure 10. View 1 (Frame 15 of Whale sharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method; (d) H. Liu et al. (e) M. R. Prabowo et al. (f), (g) Proposed ST-MRF(1st model, finer scale & coarser scale) (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale) (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

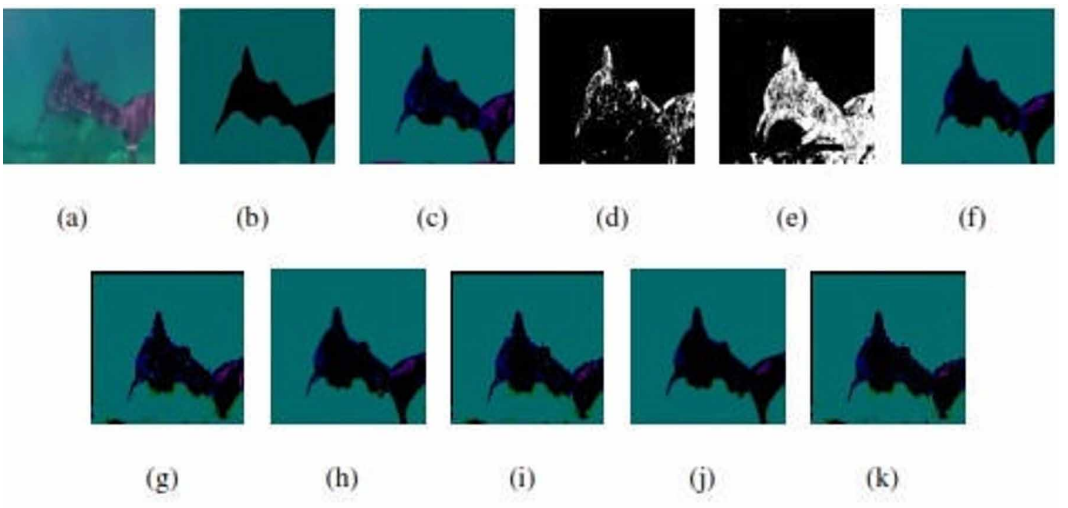


Figure 11. View 2 (Frame 16 of Whalesharks in Philippines southern Leyte, Underwater video): (a) Original image (b) Ground Truth; Segmented results using: (c) Stolkin et al.'s method (d) H. Liu et. al (e) M. R. Prabowo et al. (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale) (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale) (j),(k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

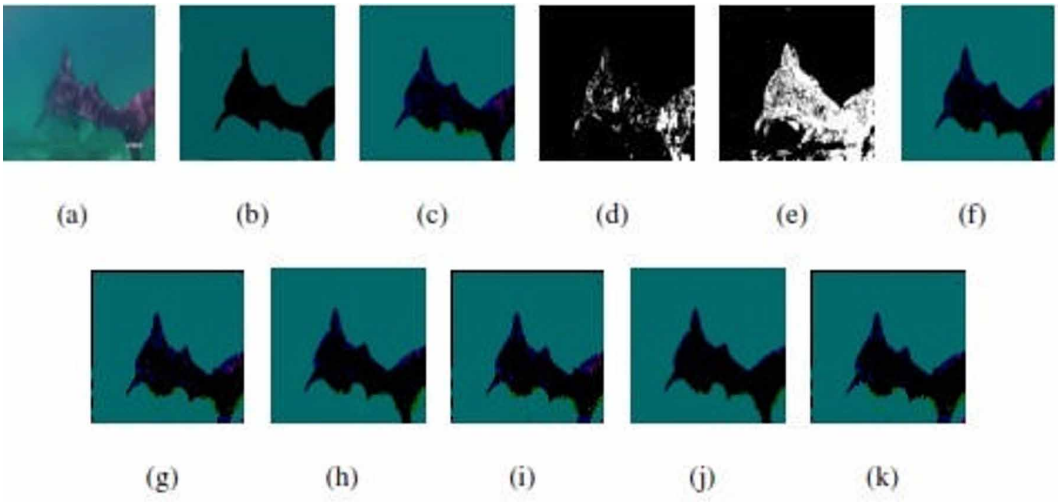
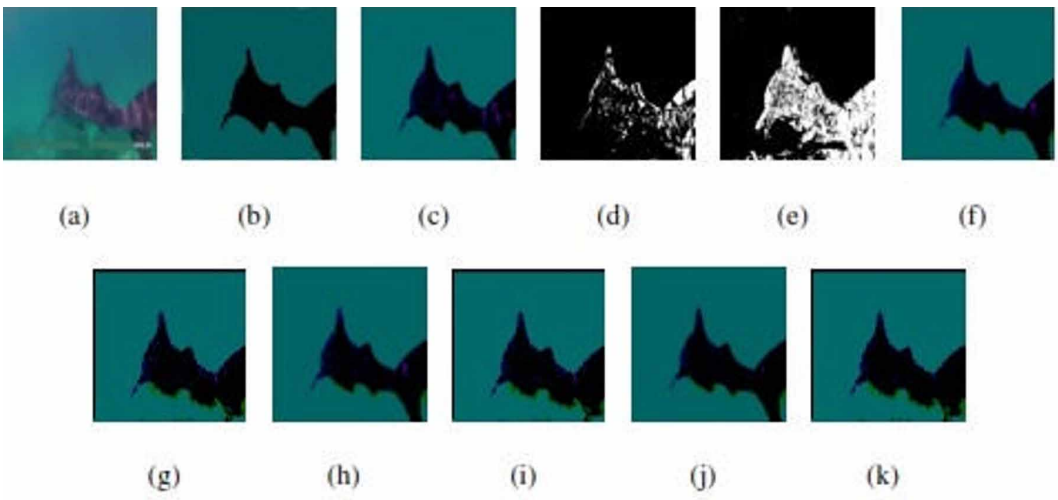


Figure 12. View 3(Frame 17 of Whalesharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al. (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale) (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale) (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).



for 3rd model is the best among all the three models. This notion has been reflected in the percentage of misclassification error presented in Table 6.

Figure 13. View 4 (Frame 18 of whale sharks in Philippines southern Leyte, Underwater video): (a) Original image (b) Ground Truth; Segmented results using: (c) Stolkin et al.'s method (d)H.Liu et al.; (e) M. R. Prabowo et al.; (f), (g)Proposed ST-MRF(1st model, finer scale & coarser scale) (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

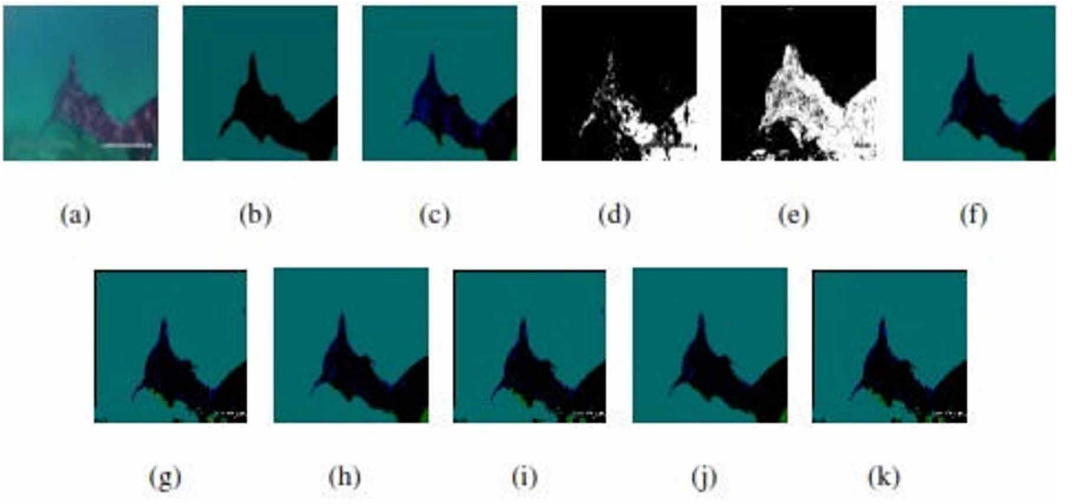
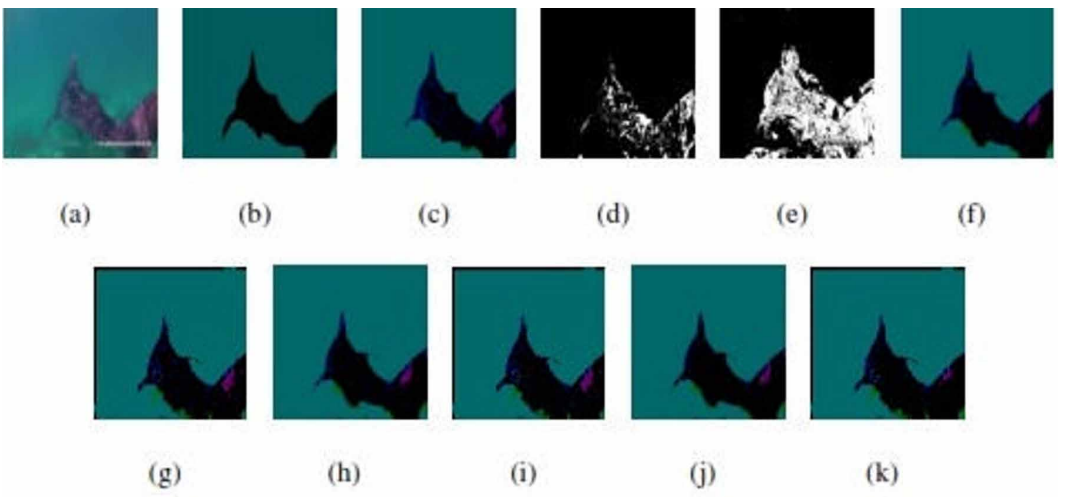


Figure 14. View 5 (Frame 19 of whale sharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF(1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).



8. QUANTITATIVE MEASURES

The accuracy of the segmented frames has been measured by the four quantitative measures such as: (i) Percentage of Misclassification Error, (ii) Dice Coefficient, (iii) Boundary Hamming Distance, (iv) Precision and Recall. The Percentage of Misclassification is defined as

Figure 15. View 6 (Frame 20 of whale sharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

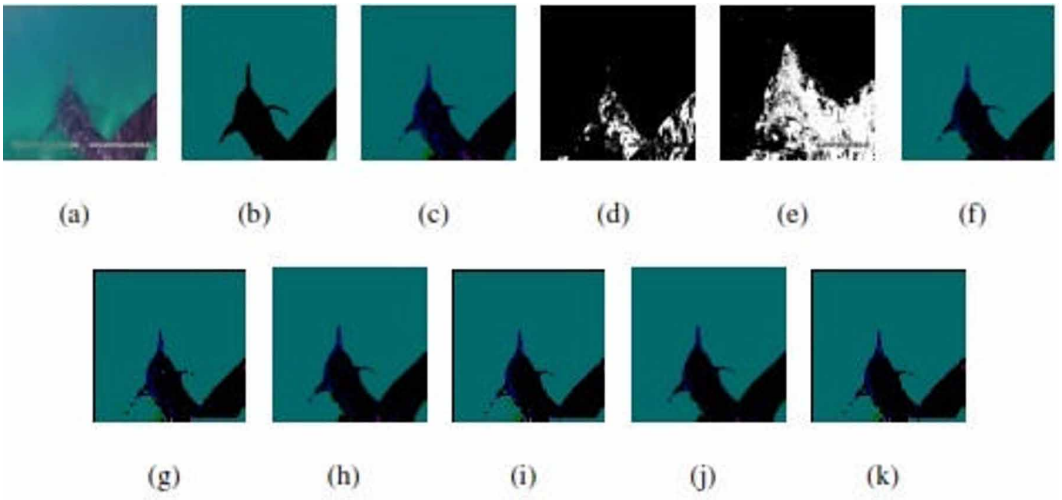
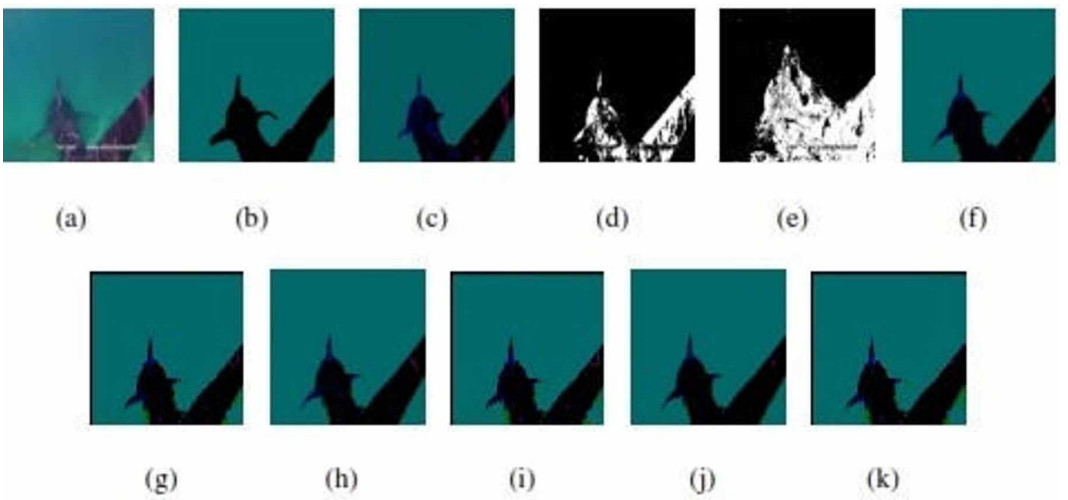


Figure 16. View 7 (Frame 21 of whale sharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF(1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).



$$Percentage\ of\ Misclassification\ Error\ (PME) = \frac{no.ofmisclassifiedpixels}{Totalnumberofpixels} \times 100$$

As observed from Table 6, the PME for the 3rd model have been found to be minimum for both fine and coarse scales. This indicates that segmentation at coarse scale is also acceptable. Figure

Figure 17. View 8 (Frame 22 of whale sharks in Philippines southern Leyte, Underwater video): (a) Original image; (b) Ground Truth; Segmented results using; (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

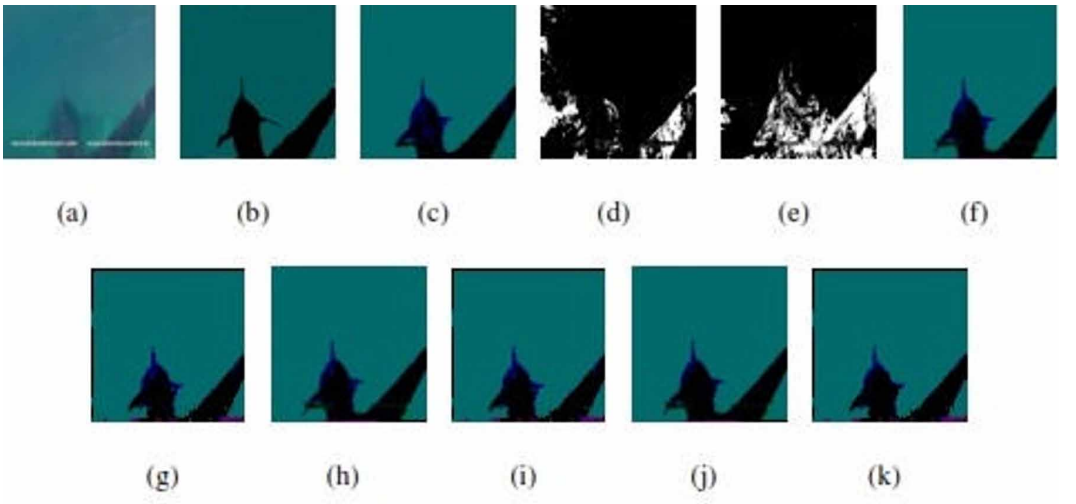
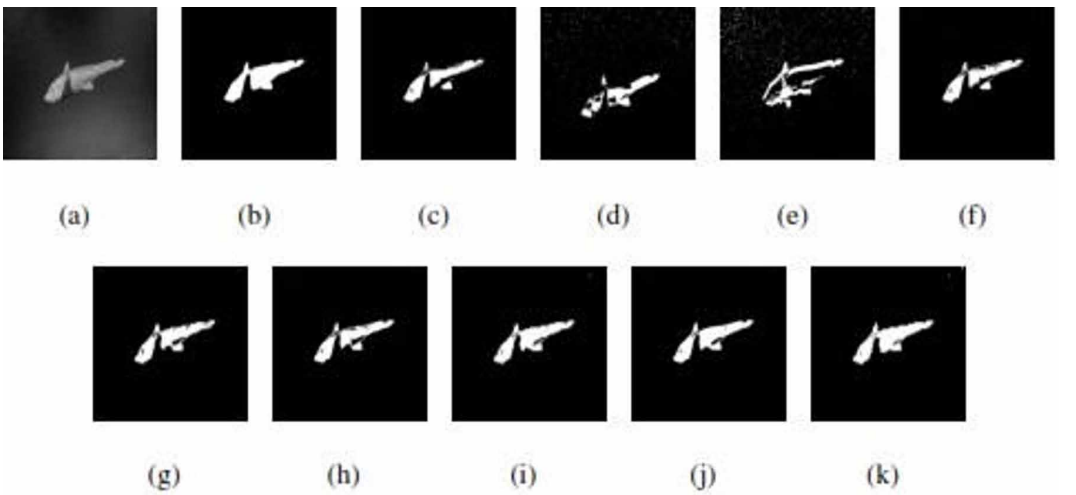


Figure 18. View 4 (Frame 32 of Creepy chimera/Nautilus live video): (a) Original image; (b) Ground Truth; Segmented results using: (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).

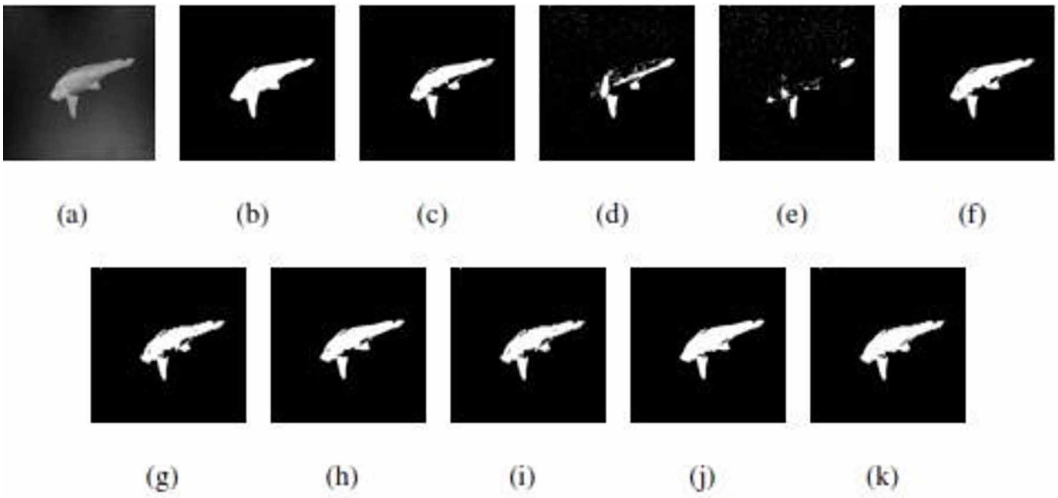


10 - Figure 17 present the segmented results for all the three models and all viewpoints. Further, it may be observed that for a given view, the PME decreases as we move from first to third model.

The second quantitative measure considered is the Dice Coefficient which is defined as

$$DC = \frac{2 \times \left| S_F \cap \mathrm{GT}_F \right|}{\left| S \right| + \left| GT \right|}$$

Figure 19. View 5 (Frame 33 of Creepy chimera/Nautilus live video): (a) Original image; (b) Ground Truth; Segmented results using: (c) Stolkin et al.'s method; (d) H. Liu et al.; (e) M. R. Prabowo et al.; (f), (g) Proposed ST-MRF (1st model, finer scale & coarser scale); (h), (i) Proposed ST-MRF (2nd model, finer scale & coarser scale); (j), (k) Proposed ST-MRF (3rd model, finer scale & coarser scale).



where, S denotes the segmented image, GT denotes the ground truth, FG and BG corresponds to foreground and background respectively.

For accurate segmentation, the Dice Coefficient values should be close to unity and in an ideal case unity. As observed from Table 7, the values corresponding to the third model at fine scale are close to unity. Further, as expected, the performance of the third model at fine scale is superior to that of the coarse scale. It is also observed that the performance of the 2nd model is superior to the first model.

The third quantitative measure Boundary Hamming Distance (BHD) is defined as

$$BHD = 1 - \frac{\left\{\left|S_B \cap GT_F\right|_{BOUNDARY} + \left|S_F \cap GT_B\right|_{BOUNDARY}\right\}}{\left|BOUNDARY\right|}$$

As observed from Table 8, the boundary hamming distance for the third model at finer scale is better than that of the coarse scale. This observation is made for all the views. The Boundary Hamming Distances are close to unity, but the distances obtained for fine scale images are higher than those of coarse scales. This is because of the blurred boundary in case of coarse scale and sharp boundaries at fine scale.

The next quantitative measures considered are Precision and Recall and are defined as

$$Precision\left(Pr\right) = \frac{TP}{TP + FP}.$$

$$Recall\left(\mathrm{Re}\right) = \frac{TP}{TP + FN}$$

The Precision and Recall values are presented in Table 9 Table 10. As observed, the Precision and Recall values are higher in case of third model at fine scale than those of coarse scale. It is

Table 6. Percentage of misclassification error at fine and coarse scale

Fine Scale							Coarse Scale		
Original views	**Stolkin's Model**	**Liu et al.**	**Prabowo et. al.**	**1st Model**	**2nd Model**	**3rd Model**	**1st Model**	**2nd Model**	**3rd Model**
Whale sharks in Philippines southern Leyte, Underwater video									
View1	15.54	8.51	7.82	6.112	6.066	5.354	6.731	6.069	5.352
View2	6.71	9.0	6.33	1.496	1.466	0.925	1.499	1.469	0.926
View3	5.95	8.92	6.94	0.831	0.825	0.503	0.828	0.822	0.504
View4	6.34	7.34	7.86	6.221	6.001	5.002	6.218	6.002	5.000
View5	7.33	9.15	7.61	4.181	4.001	3.882	4.921	4.008	3.880
View6	6.19	7.76	9.69	0.613	0.560	0.320	0.610	0.558	0.318
View7	10.1	5.78	9.47	8.071	7.756	6.863	8.068	7.753	6.864
View8	6.69	8.8	8.51	0.461	0.449	0.293	0.465	0.445	0.290
Creepy chimera/Nautilus live video									
View4	1.46	8.68	5.3	1.24	1.04	0.62	2.44	1.79	0.85
View5	1.20	4.2	6.6	0.82	0.74	0.41	1.88	1.76	1.05

further observed that the Precision and Recall values for all the models at fine scale are higher than those of coarse scale. However, as expected the results obtained for coarse scale are close to those at fine scale. The Recall values are high but not close to unity as in the case of Precision values. The corresponding values of all the three existing algorithms are lower. Hence, the algorithm could be successfully tested with two underwater video data sets. The performance of the proposed algorithms is superior to that of Stolkin's EMRF, H. Liu et. al, M. R. Prabowo et. al algorithms.

9. CONCLUSION

This research has proposed new scheme to address the issue when both object and the camera are moving in the underwater environment. Since this problem is a challenging issue in a real-world scenario, which motivated us to consider in this research work. This has been viewed as an incomplete data problem and the problem has been solved in EM framework. In the proposed framework apriori knowledge of the camera model parameters are assumed to be available. Three Spatio Temporal MRF models have been proposed as a priori models which are used to estimate the labels in the MAP framework. In the M step, new features are computed to estimate the camera model's intrinsic and extrinsic parameters. The three proposed algorithms could successfully be tested with frames derived from two video sequences. The results obtained for different frames correspond to fine scale operations and were found to be superior to that of existing three algorithms. The efficacy of the algorithms has also been tested at coarse resolution. The coarse resolution frames could be successfully used to detect the objects and it has been observed that the results are acceptable based on the different quantitative measures. But in case of coarse scale, the computational time has reduced substantially thus, making it a feasible candidate for real time applications.

Table 7. Dice coefficients at fine and coarse scale

Fine Scale							Coarse Scale		
Original views	**Stolkin's Model**	**Liu et al.**	**Prabowo et al.**	**1st Model**	**2nd Model**	**3rd Model**	**1st Model**	**2nd Model**	**3rd Model**
Whale sharks in Philippines southern Leyte, Underwater video									
View1	.664	.847	.860	0.919	0.922	0.931	0.908	0.920	0.935
View2	.854	.838	.887	0.926	0.935	0.943	0.924	0.931	0.941
View3	.870	.840	.876	0.936	0.941	0.951	0.932	0.945	0.950
View4	.863	.868	.860	0.891	0.893	0.899	0.894	0.890	0.900
View5	.842	.837	.864	0.924	0.927	0.933	0.893	0.924	0.935
View6	.866	.862	.827	0.939	0.941	0.945	0.944	0.945	0.947
View7	.855	.897	.831	0.852	0.854	0.859	0.858	0.859	0.859
View8	.782	.841	.848	0.945	0.947	0.950	0.943	0.949	0.951
Creepy chimera/Nautilus live video									
View4	.885	.913	.946	.977	.979	.984	.966	.970	.973
View5	.887	.957	.933	.981	.982	.985	.967	.968	.973

Table 8. Boundary hamming distance at fine and coarse scale

Fine Scale							Coarse Scale		
Original views	**Stolkin's Model**	**Liu et al.**	**Prabowo et al.**	**1st Model**	**2nd Model**	**3rd Model**	**1st Model**	**2nd Model**	**3rd Model**
Whale sharks in Philippines southern Leyte, Underwater video									
View1	.925	.655	.654	0.875	0.876	0.865	0.870	0.872	0.835
View2	.911	.611	.678	0.910	0.908	0.909	906	0.905	0.906
View3	952	.709	.619	0.946	0.947	0.947	0.943	0.945	0.948
View4	.943	.716	.549	0.948	0.948	0.948	0.944	0.942	0.945
View5	.901	.699	.618	0.867	0.867	0.868	0.860	0.863	0.863
View6	.905	.618	.517	0.926	0.926	0.928	0.923	0.922	0.923
View7	.918	.761	.597	0.940	0.940	0.940	0.933	0.933	0.932
View8	.900	.674	.518	0.889	0.884	0.881	0.885	0.882	0.882
Creepy chimera/Nautilus live video									
View4	.842	.454	.695	.874	.883	.915	.880	.882	.882
View5	.881	.753	.631	.902	.909	.929	.901	.905	.920

Table 9. Precision at Fine and Coarse Scale

Fine Scale							Coarse Scale		
Original views	**Stolkin's Model**	**Liu et al.**	**Prabowo et al.**	**1st Model**	**2nd Model**	**3rd Model**	**1st Model**	**2nd Model**	**3rd Model**
Whale sharks in Philippines southern Leyte, Underwater video									
View1	.993	.932	.640	0.995	0.995	0.995	0.978	0.978	0.979
View2	.996	.918	.713	0.997	0.998	0.998	0.976	0.978	0.978
View3	993	.931	.700	0.993	0.993	0.994	0.829	0.829	0.827
View4	.993	.851	.627	0.997	0.997	0.997	0.801	0.804	0.805
View5	.990	.905	.649	0.993	0.994	0.995	0.893	0.895	0.896
View6	.986	.969	.559	0.989	0.995	0.995	0.964	0.966	0.967
View7	.984	.972	.562	0.995	0.989	0.990	0.871	0.874	0.874
View8	.993	.513	.528	0.995	0.995	0.995	0.972	0.976	0.977
Creepy chimera/Nautilus live video									
View4	.993	.181	.580	.996	.996	.997	.996	.996	.998
View5	.995	.879	.588	.997	.998	.998	.997	.998	.998

Table 10. Recall at Fine and Coarse Scale

Fine Scale					Coarse Scale				
Original views	**Stolkin's Model**	**Liu et al.**	**Prabowo et al.**	**1st Model**	**2nd Model**	**3rd Model**	**1st Model**	**2nd Model**	**3rd Model**
Whale sharks in Philippines southern Leyte, Underwater video									
View1	.974	.262	.707	0.976	0.976	0.974	0.905	0.907	0.900
View2	.982	.261	.782	0.985	0.984	0.984	0.913	0.911	0.911
View3	994	.280	.748	0.995	0.995	0.995	0.893	0.893	0.892
View4	.990	.446	.801	0.992	0.991	0.992	0.876	0.874	0.874
View5	.971	.235	.739	0.976	0.977	0.977	0.879	0.875	0.874
View6	.973	.340	.742	0.976	0.976	0.974	0.947	0.946	0.944
View7	.998	.515	.803	0.998	0.998	0.998	0.880	0.876	0.875
View8	.988	.258	.512	0.990	0.989	0.988	0.937	0.938	0.938
Creepy chimera/Nautilus live video									
View4	.729	.114	.482	.773	.813	.899	.945	.981	.998
View5	.834	.456	.197	.890	.901	.959	.959	.968	.994

Table 11. Execution time (ET) of Stolkin's Model and 1st Model (Fine and Coarse Scale) (Whalesharks in Philippines southern Leyte, Underwater video)

Original views	ET of Stolkin's Model (sec) (480×270)	ET of Fine Scale (sec) (480×270)	ET of Coarse Scale (sec) (120×67)
View1	142	142	8
View2	141	143	9
View3	143	143	8
View4	143	145	10
View5	144	145	10
View6	145	147	11
View7	141	143	9
View8	140	143	9

10. REFERENCES

Ancuti, C., Ancuti, C. O., Haber, T., & Bekaert, P. (2012). Enhancing underwater images and videos by fusion. *Paper presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Press. doi:10.1109/CVPR.2012.6247661

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B. Methodological*, *48*(3), 259–279. doi:10.1111/j.2517-6161.1986.tb01412.x

Boudhane, M., & Nsiri, B. (2016). Underwater image processing method for fish localization and detection in submarine environment. *Journal of Visual Communication and Image Representation*, *39*, 226–238. doi:10.1016/j.jvcir.2016.05.017

Chandan, K. R., & Bala, R. (2009). Theory and Practice of Expectation Maximization (EM) Algorithm. In W. John (Ed.), *Encyclopedia of Data Warehousing and Mining* (2nd ed., pp. 1966–1973). Hershey, PA: IGI Global.

Cho, S.-H., Jung, H.-K., Lee, H., Rim, H., & Lee, S. K. (2016). Real-time underwater object detection based on DC resistivity method. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(11), 6833–6842. doi:10.1109/TGRS.2016.2591619

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x

Emberton, S., Chittka, L., & Cavallaro, A. (2018). Underwater image and video dehazing with pure haze region segmentation. *Computer Vision and Image Understanding*, *168*, 145–156. doi:10.1016/j.cviu.2017.08.003

Geman, S., & Geman, D. (1987). *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In Readings in computer vision* (pp. 564–584). Elsevier.

Halder, K. K., Tahtali, M., & Anavatti, S. G. (2016). Moving object detection and tracking in videos through turbulent medium. *Journal of Modern Optics*, *63*(11), 1015–1021. doi:10.1080/09500340.2015.1117665

Heikkila, J., & Silven, O. (1997). A four-step camera calibration procedure with implicit image correction. *Paper presented at the CVPR*. Academic Press. doi:10.1109/CVPR.1997.609468

Hossain, E., Alam, S. S., Ali, A. A., & Amin, M. A. (2016). Fish activity tracking and species identification in underwater video. *Paper presented at the 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. Academic Press. doi:10.1109/ICIEV.2016.7760189

Kang, L., Wu, L., Wei, Y., Lao, S., & Yang, Y.-H. (2017). Two-view underwater 3D reconstruction for cameras with unknown poses under flat refractive interfaces. *Pattern Recognition*, *69*, 251–269. doi:10.1016/j.patcog.2017.04.006

Karasaridis, A., & Simoncelli, E. (1996). A filter design technique for steerable pyramid image transforms. *Paper presented at the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE Press. doi:10.1109/ICASSP.1996.547763

Li, S. Z. (1994). Markov random field models in computer vision. *Paper presented at the European conference on computer vision*. Academic Press. doi:10.1007/BFb0028368

Liu, H., Dai, J., Wang, R., Zheng, H., & Zheng, B. (2016, April 10-13). Combining background subtraction and three-frame difference to detect moving object from underwater video. *Paper presented at the OCEANS 2016*. Academic Press.

Mohapatra, S. K., Mahapatra, S. K., Mahapatra, S., & Swain, B. R. (2015). Simulation based algorithm for tracking fish population in unconstrained underwater. *Paper presented at the 2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE)*. Academic Press. doi:10.1109/ICMOCE.2015.7489767

Negrea, C., Thompson, D. E., Juhnke, S. D., Fryer, D. S., & Loge, F. J. (2014). Automated detection and tracking of adult Pacific Lampreys in underwater video collected at Snake and Columbia River fishways. *North American Journal of Fisheries Management*, *34*(1), 111–118. doi:10.1080/02755947.2013.849634

Panda, S., & Nanda, P. K. (2015). Segmentation of underwater video objects using extended Markov random field model. *Paper presented at the 2015 IEEE Underwater Technology (UT)*. IEEE Press. doi:10.1109/UT.2015.7108255

Prabowo, M., Hudayani, N., Purwiyanti, S., Sulistiyanti, S., & Setyawan, F. (2017). A moving objects detection in underwater video using subtraction of the background model. *Paper presented at the 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. Academic Press. doi:10.1109/EECSI.2017.8239148

Qiao, Y., Tang, Y., & Li, J. (2013). Improved Harris sub-pixel corner detection algorithm for chessboard image. *Paper presented at the 2013 2nd International Conference on Measurement, Information and Control*. Academic Press.

Silvatti, A. P., Cerveri, P., Telles, T., Dias, F. A., Baroni, G., & Barros, R. M. (2013). Quantitative underwater 3D motion analysis using submerged video cameras: Accuracy analysis and trajectory reconstruction. *Computer Methods in Biomechanics and Biomedical Engineering*, *16*(11), 1240–1248. doi:10.1080/10255842.2012.664637 PMID:22435960

Stolkin, R., Greig, A., Hodgetts, M., & Gilby, J. (2008). An EM/E-MRF algorithm for adaptive model based tracking in extremely poor visibility. *Image and Vision Computing*, *26*(4), 480–495. doi:10.1016/j.imavis.2007.06.008

Stolkin, R., Hodgetts, M., Greig, A., & Gilby, J. (2007). Extended Markov random fields for predictive image segmentation *Advances In Pattern Recognition*, 208–214.

Walther, D., Edgington, D. R., & Koch, C. (2004). Detection and tracking of objects in underwater video. *Paper presented at the Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004*. IEEE Press. doi:10.1109/CVPR.2004.1315079

Zhang, D., Kopanas, G., Desai, C., Chai, S., & Piacentino, M. (2016). Unsupervised underwater fish detection fusing flow and objectiveness. *Paper presented at the 2016 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE Press. doi:10.1109/WACVW.2016.7470121

Zhang, W., Liang, J., Ju, H., Ren, L., Qu, E., & Wu, Z. (2017). Study of visibility enhancement of hazy images based on dark channel prior in polarimetric imaging. *Optik-International Journal for Light and Electron Optics*, *130*, 123–130. doi:10.1016/j.ijleo.2016.11.047

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330-1334.

Zhou, F., Cui, Y., Peng, B., & Wang, Y. (2012). A novel optimization method of camera parameters used for vision measurement. *Optics & Laser Technology*, *44*(6), 1840–1849. doi:10.1016/j.optlastec.2012.01.023

Susmita Panda is an Assistant Professor in the Department of Electronics and Communication, Siksha 'O' Anusandhan, Deemed to be University, Bhubaneswar, Odisha, India. She received her master's degree in 2009 from the same University. Currently, she is pursuing her Ph.D. She is member of IEEE Geoscience and Remote Sensing Society. Her research interests are image processing, video object detection, and computer vision.

Pradipta Kumar Nanda obtained his Ph.D from obtained his PhD from IIT Bombay in 1996 in the area of Computer Vision. He has served more than 21 years at NIT Rourkela with the last position as professor and Head of the Department. Currently, he is the Dean (Research & Development) of Siksha o Anusandhan, Deemed to be University, Bhubaneswar, Odisha, India. He has published 82 papers in various Journals and Conference proceedings and has authored 4 research books. He was an academic visitor to School of Computing, University of Leeds, UK in March 2006 and delivered lectures to their faculties and students. He is a Fellow of IETE, India and Senior Member of IEEE, USA, Member IET and Life Member of ISTE, India. His research interests are bio-medical image analysis, video object detection and tracking, computer vision, soft computing and their applications, and ad-hoc wireless sensor networks.