A Change Information Based Fast Algorithm for Video Object Detection and Tracking

Badri Narayan Subudhi, Pradipta Kumar Nanda, *Member, IEEE,* and Ashish Ghosh, *Member, IEEE*

***Abstract*—In this paper, we present a novel algorithm for moving object detection and tracking. The proposed algorithm includes two schemes: one for spatio-temporal spatial segmentation and the other for temporal segmentation. A combination of these schemes is used to identify moving objects and to track them. A compound Markov random field (MRF) model is used as the prior image attribute model, which takes care of the spatial distribution of color, temporal color coherence and edge map in the temporal frames to obtain a spatio-temporal spatial segmentation. In this scheme, segmentation is considered as a pixel labeling problem and is solved using the maximum *a posteriori* probability (MAP) estimation technique. The MRF-MAP framework is computation intensive due to random initialization. To reduce this burden, we propose a change information based heuristic initialization technique. The scheme requires an initially segmented frame. For initial frame segmentation, compound MRF model is used to model attributes and MAP estimate is obtained by a hybrid algorithm [combination of both simulated annealing (SA) and iterative conditional mode (ICM)] that converges fast. For temporal segmentation, instead of using a gray level difference based change detection mask (CDM), we propose a CDM based on label difference of two frames. The proposed scheme resulted in less effect of silhouette. Further, a combination of both spatial and temporal segmentation process is used to detect the moving objects. Results of the proposed spatial segmentation approach are compared with those of JSEG method, and *edgeless* and *edgebased* approaches of segmentation. It is noticed that the proposed approach provides a better spatial segmentation compared to the other three methods.**

***Index Terms*—Image edge analysis, image motion analysis, image segmentation, MAP estimation, modeling, object detection, simulated annealing, tracking.**

Nomenclature

MRF	Markov random field.
SA	Simulated annealing.
ICM	Iterative conditional mode.
MAP	Maximum *a posteriori* probability.
CDM	Change detection mask.
VOP	Video object plane.
JSEG	Joint segmentation scheme.
DGA	Distributed genetic algorithm.
RGB	Red, green and blue.
det	Determinant.
$no.$	Number.
t	tth time instant.
$t-d$	$(t-d)$th time instant.
y	Observed video sequence.
y_t	Observed image frame at time t.
s	Site.
y_{st}	A site s of frame y_t.
x	Segmentation of y.
X_t	Markov random field.
x_t	Realization of X_t, segmented version of y_t.
$\eta_{s,t}$	Neighborhood of (s) in spatial direction of the tth frame.
$V_{sc}(x_t)$	Clique potential function in the spatial domain.
$V_{tec}(x_t)$	Clique potential in the temporal domain.
$V_{teec}(x_t)$	Clique potential in the temporal domain incorporating edge features.
α, β and γ	Clique potential parameters.
$\hat{x}_t$	Estimated label or MAP estimate.
θ	Parameter vector associated with x_t.
$U(x_t)$	Energy of realization x_t.
N	Gaussian process.
k	Covariance matrix.
n	Realization of the Gaussian process $N(\mu, \sigma)$.
μ	Mean of Gaussian process.
σ^2	Variance of Gaussian process.
$y_{(t+d)_{\|y_{t+d}-y_t\|}}$	Changed region corresponding to y_{t+d} frame.
x_{tt}	Changed region obtained in $y_{(t+d)_{\|y_{t+d}-y_t\|}}$.
$x_{(t+d)i}$	Initialization of the $(t+d)$th frame.
R	VOP image matrix.
r_m	Value of VOP at location m.
$(\hat{u}_{n_c}, \hat{v}_{n_c})$	Centroid of the moving object.
$Gain$	Time gain.
f	Number of features.

I. Introduction

Detection and tracking of moving objects from a video scene is a challenging task in video processing and computer vision [1]–[4]. It has wide applications such as video surveillance, event detection, activity recognition, activity based human recognition, fault diagnosis, anomaly

Manuscript received May 8, 2009; revised October 5, 2009 and July 9, 2010; accepted January 14, 2011. Date of publication March 28, 2011; date of current version July 7, 2011. This paper was recommended by Associate Editor R. Green.

B. N. Subudhi and A. Ghosh are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: subudhi.badri@gmail.com; ash@isical.ac.in).

P. K. Nanda is with the Department of Electronics and Telecommunication Engineering, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar 751030, India (e-mail: pknanda.nitrkl@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

detection, robotics, autonomous navigation, dynamic scene analysis, path detection, and others [1]–[4]. Moving object detection in a video is the process of identifying different object regions which are moving with respect to the background. More specifically, moving object detection in a video is the process of identifying those objects in the video whose movements will create a dynamic variation in the scene [2]. This can be achieved by two different ways: 1) motion detection/change detection, and 2) motion estimation [2]. Change or motion detection is the process of identifying changed and unchanged regions from the extracted video image frames when the camera is fixed and the objects are moving. For motion estimation, we compute the motion vectors to estimate the positions of the moving objects from frame to frame. In case of motion estimation, both the objects and the camera may move [2]. After detecting the moving objects from the image frames, it is required to track them. Tracking of a moving object from a video sequence helps in finding the velocity, acceleration, and position of it at different instants of time. In visual surveillance, sometimes it may be required to obtain the speed/velocity of a moving vehicle so as to keep an eye on the movement of a particular vehicle [16].

Moving object detection by the process of motion/change detection is again restricted by the requirement of a reference frame (where the object is not present). This can be accomplished by the use of intensity difference based motion detection algorithm [4] (where objects may move slow or fast). In the absence of a reference frame, if there is a substantial amount of movement of an object from one frame to another, the object can be tracked exactly by generating a reference frame [4]. However, for those cases where the reference frame is not available and: 1) the objects in the scene do not have a substantial amount of movement from frame to frame, or 2) the objects in a given scene move and stop for some time and move further, identification of moving objects becomes difficult with temporal segmentation [1]–[4].

A robust video image segmentation algorithm is essential to solve these problems. Watershed algorithm (a region based approach) [1], [2], [4] is a famous approach in this context. A computationally efficient watershed based spatial segmentation approach was proposed by Salember *et al.* [5]. They had used spatial segmentation and temporal segmentation to detect object boundaries. However, this method produced oversegmented results and hence could not detect the objects satisfactorily.

Different stochastic model based approaches [8] are available in the literature and they provide better results. MRF model, because of its attribute to model spatial dependency, is proved to be a better model for image segmentation [10]. MRF models [9]–[23] and Hidden MRF models [27], [28] have also been used for moving object detection for the last two decades. Since in a video, spatial and temporal coherence is there, MRF model is shown to be a better resilience. An early work on MRF based object detection scheme was proposed by Hinds *et al.* [11]. In order to obtain a smooth transition of segmentation results from frame to frame, temporal constraints was introduced. They had adhered to a multi-resolution approach to reduce the computational burden. A similar approach where MRF model had been used to obtain a 3-D spatio-temporal segmentation was proposed by Wu *et al.* [12]. In this paper, region growing approach along with contour relaxation was applied to obtain accurate boundaries of objects. In the approach proposed by Babacan *et al.* [19], the video sequences were modeled as MRF, a spatial direction MRF model was used for spatial segmentation. Previous frame segmentation result acted as the precauser for the next frame segmentation. This method could overcome the fragmentation caused by the spatio-temporal framework of Hinds *et al.* [11]. A similar approach where the changes in temporal direction had been modeled by a mixture of Gaussian MRFs was also proposed by Babacan *et al.* [20] for detection of moving objects. In this case, the MAP estimate was obtained by ICM. They proposed a scheme of background modeling that exploited both spatial and temporal dependency to improve the quality of segmentation of both indoor and outdoor surveillance videos. However, all these methods [19], [20] are constrained to assume the availability of the reference frame. These methods fail to segment the targets in the absence of reference frame and also fail when temporal changes in between the frames are not substantial.

All the MRF model based approaches discussed so far were used for video object detection along with spatial segmentation, whereas combination of spatial segmentation along with temporal segmentation proved to be a better choice of detecting moving objects [13], [14], [17]. In these methods MRFs have been used to model video image frames and spatial segmentation problem has been formulated in spatio-temporal framework. The spatio-temporal spatial segmentation thus obtained is combined with the results of temporal segmentation to detect the moving objects. Kim *et al.* [13] proposed a video object detection scheme where each video sequence was modeled with a MRF, and the MAP estimate was obtained using DGA. The unstable chromosomes found during the evolution from frame to frame were regarded as moving objects. A similar approach was also proposed by Hwang *et al.* [14]. In this scheme, spatial segmentation was obtained using MRF model and DGA was used to obtain the MAP estimate. The temporal segmentation was obtained by direct combination of VOP of the previous frame with the CDM of the current frame. The objects from the previous frame were assumed to be present in the current frame also and lead it to an error in the detection of moving objects in the current frame correctly. This gave an effect of silhouette. The effect of silhouette is found to be less in a recently proposed moving object detection technique of Kim *et al.* [17]. They had extended the video segmentation scheme proposed by Hwang *et al.* [14] where an evolutionary probability was considered to update the crossover and mutation rate through evolution in DGA. Thereafter for temporal segmentation of current frame, the CDM was updated with the label information of the current and the previous frames. Combination of both spatio-temporal spatial segmentation and temporal segmentation was performed to obtain the VOP, that gives an accurate shape of moving objects, with less effect of silhouette.

A region labeling approach that uses MRF model with motion estimation was used by Tsaig *et al.* [15] to obtain the VOPs. Here, to obtain an initial partition of the considered frame, watershed algorithm was used. Recently a region based object detection algorithm was proposed by Huang *et al.* [21] to obtain a set of motion coherent regions. They had also used MRFs for spatial segmentation and integrated the spatial as well as temporal sequences to obtain the moving objects.

An adaptive thresholding based background and foreground separation scheme for target detection was proposed by Kim *et al.* [29]. The intensity distribution of the video sequence had been modeled by Gaussian distribution and the parameters had been estimated using an auto regressive model. The objects and background were classified and thereafter the objects were tracked by checking the movement of the centroids of the identified objects. This yielded quite satisfactory results for video surveillance.

In this paper, we propose a compound MRF model [25] based scheme that detects moving objects with less computational burden. This method is able to track moving objects in the absence of any reference frame, and when objects are moving very slowly or do not have much movements. The proposed scheme is a combination of both spatio-temporal spatial segmentation and temporal segmentation. Here, we obtain spatio-temporal spatial segmentation first for a given initial frame by *edgebased* compound MRF Model. Thereafter, for subsequent frames, segmentation is obtained by adding some change information of these frames with initial frame segmentation result. In the *edgebased* compound MRF model [25] of segmentation, a compound MRF model is used that takes care of the spatial distribution of the current frame, temporal frames and edge maps in the temporal direction. This problem is formulated using MAP estimation principle. For the initial image frame, the MAP estimate is obtained using a hybrid algorithm. For subsequent frames, original pixels corresponding to the changed regions (changes obtained between current and previously considered frames) of the current frame are super-imposed on previously available segmented frame to obtain a heuristic initialization. Subsequent frames are modeled with compound MRFs and the MAP estimate is obtained by ICM algorithm starting from this initialization. This spatio-temporal spatial segmentation combined with temporal segmentation yields the VOP and hence can detect moving objects. For temporal segmentation, we used a label difference CDM instead of a gray level difference CDM. Moment of inertia based tracking strategy is used to track moving objects from a given video sequence.

The results obtained by the proposed spatio-temporal spatial segmentation method are compared with those of JSEG [24], *edgeless* [25] and *edgebased* [25] methods of segmentation and is found to be better. Computational time requirement for the proposed method is less compared to *edgeless* and *edgebased* approaches. Similarly the results obtained for VOP by the label frame difference CDM is compared with those of CDM with a gray level difference, and it is found that the VOP with label frame difference CDM approach gives better results.

The organization of this paper is as follows. In Section II, algorithm for detecting objects is narrated with the help of a block diagram. In Section III, spatial segmentation method using spatio-temporal framework is presented where initial frame segmentation and change information based subsequent frame segmentation schemes are described along with spatio-temporal MRF based image modeling. In Section IV, temporal segmentation based on CDM is elaborated. In Section V, VOP generation and tracking process by centroid based method is discussed. Section VI provides simulation results and analysis. Our conclusion is presented in Section VII.

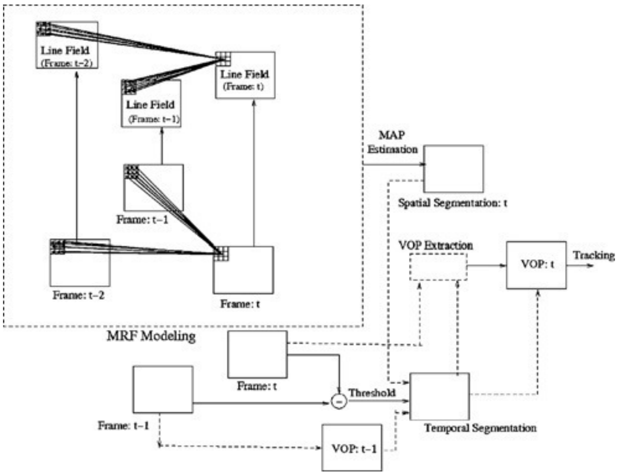

Fig. 1. Block diagram of the proposed scheme.

II. Proposed Algorithm for Object Detection

A block diagrammatic representation of the proposed scheme is given in Fig. 1. Here we use two types of segmentation schemes: one is a spatio-temporal spatial segmentation and the other is a temporal segmentation. Spatial segmentation helps in determining the boundary of the regions in the scene accurately, and temporal segmentation helps in determining the foreground and the background parts of it.

The spatial segmentation task is considered in spatio-temporal framework. Here the attributes like color or gray value in the spatial direction, color or gray value in the temporal direction and edge map/line field both in spatial and temporal directions are modeled with MRFs. RGB color model is used. The edge map considered is obtained by considering a 3×3 Laplacian window. In order to speed up the algorithm, initial image frame is segmented with *edgebased* spatio-temporal modeling and a hybrid algorithm (hybrid of both SA [22] and ICM [23]) is used for MAP estimation. For subsequent frames, a change information based algorithm is proposed with less computational burden.

For temporal segmentation, a CDM is obtained by taking the difference between two consecutive frames, where information from the previous frame is fed back and the label of the current spatial segmentation result is used to modify the CDM. The modified CDM itself represents a binary mask of foreground and background region where VOP is extracted by superimposing the original pixels of the current frame on the foreground part of the temporal segmentation.

A schematic representation of the whole process is shown in Fig. 1. Here we assume the frame instants to be the same as the time instants. Frame t represents the observed image frame at tth instant of time. We model the tth frame with its second order neighbors both in spatial and temporal directions. For temporal direction modeling, we have considered two temporal frames at $(t-1)$th and $(t-2)$th instants. Similarly edge/line field of tth frame is modeled with its neighbors in temporal direction at $(t-1)$th and $(t-2)$th frames. The estimated MAP of the MRF represents the spatial segmentation result of the tth frame. The whole process is performed in spatio-temporal framework, and hence is termed as spatio-temporal spatial segmentation. For temporal segmentation we have used a motion detection scheme. We obtain a difference image of two consecutive frames, i.e., the tth and the $(t-1)$th frame and is thresholded by a suitable threshold value. The thresholded image itself represents the amount of movement performed by objects in the scene from the $(t-1)$th instant to the tth instant of time. The spatial segmentation result of the tth frame, the $(t-1)$th frame, along with VOP of the $(t-1)$th frame were used to perform a temporal segmentation of the tth frame. The pixels corresponding to the object regions of the temporal segmented output are replaced by the original pixels of the tth frame to obtain the VOP of the tth frame.

We have considered a moment of inertia based scheme to find out the centroid of a detected object. Moving objects are tracked by calculating the centroid of the detected objects from frame to frame.

III. Spatio-Temporal Spatial Segmentation

In the spatio-temporal spatial segmentation scheme, we have modeled each video image frame with compound MRF model and the segmentation problem is solved using the MAP estimation principle. For initial frame segmentation, a hybrid algorithm is proposed to obtain the MAP estimate. For segmentation of other frames, changes between the frames is imposed on the previously available segmented frame so as to have an initialization to find the segmentation result of other frames. The total scheme is described in detail in the subsequent sections.

A. MRF Based Spatio-Temporal Image Modeling

Here it is assumed that the observed video sequence y is a 3-D volume consisting of spatio-temporal image frames. y_t represents a video image frame at time t and hence is a spatial entity. Each pixel in y_t is a site s denoted by y_{st}. Let Y_t represent a random field and y_t be a realization of it at time t. Thus, y_{st} denotes a spatio-temporal co-ordinate of the grid (s,t). Let x denote the segmentation of video sequence y and x_t denote the segmented version of y_t. Let us assume that X_t represents the MRF from which x_t is a realization. Similarly, pixels in the temporal direction are also modeled as MRFs. We have considered the second order MRF modeling both in spatial and in temporal directions. In order to preserve the edge features, another MRF model is considered with the linefield of the current frame x_t and the line fields of x_{t-1} and x_{t-2}. It is known that if X_t is a MRF then it satisfies the Markovianity property in spatial direction, that is

$$\begin{aligned} &P(X_{st}=x_{st} \mid X_{qt}=x_{qt}, \forall q \epsilon S, s \neq q) = \\ &\qquad P(X_{st}=x_{st} \mid X_{qt}=x_{qt}, (q,t) \epsilon \eta_{s,t}) \end{aligned}$$

where $\eta_{s,t}$ denotes the neighborhood of (s,t) and S denotes the spatial lattice of X_t. For temporal MRF, the following Markovianity property is also satisfied:

$$\begin{aligned} &P(X_{st}=x_{st} \mid X_{pq}=x_{pq}, q \neq t, p \neq s, \forall (s,t) \epsilon V) = \\ &\qquad P(X_{st}=x_{st} \mid X_{pq}=x_{pq}, (p,q) \epsilon \eta_{s,t}). \end{aligned}$$

Here V denotes the 3-D volume of the video sequence. In spatial domain, X_t represents the MRF model of x_t and hence the prior probability can be expressed as Gibb's distribution with $P(X_t) = \frac{1}{z} e^{\frac{-U(X_t)}{T}}$, where z is the partition function expressed as $z = \sum_{x_t} e^{\frac{-U(x_t)}{T}}$, $U(X_t)$ is the energy function (a function of clique potentials). We have considered the following clique potential functions for the present work:

$$V_{sc}(x_t) = \begin{cases} +\alpha, & \text{if} \quad x_{st} \neq x_{pt} \text{ and } (s,t),(p,t) \epsilon S \\ -\alpha, & \text{if} \quad x_{st} = x_{pt} \text{ and } (s,t),(p,t) \epsilon S. \end{cases}$$

Analogously in the temporal direction

$$V_{tec}(x_t) = \begin{cases} +\beta, & \text{if} \quad x_{st} \neq x_{qt} \text{ and } (s,t),(q,t-1) \epsilon S \\ -\beta, & \text{if} \quad x_{st} = x_{qt} \text{ and } (s,t),(q,t-1) \epsilon S \end{cases}$$

and for the edgemap in the temporal direction as

$$V_{teec}(x_t) = \begin{cases} +\gamma, & \text{if} \quad x_{st} \neq x_{et} \text{ and } (s,t),(e,t-1) \epsilon S \\ -\gamma, & \text{if} \quad x_{st} = x_{et} \text{ and } (s,t),(e,t-1) \epsilon S. \end{cases}$$

Here α, β and γ are the parameters associated with the clique potential function. These are $+ve$ constants and are determined on a trial and error basis.

In image modeling the clique potential function is the combination of the above three terms. Hence, the energy function is of the following form:

$$U(X_t) = \sum_{c \in C} V_{sc}(x_t) + \sum_{c \in C} V_{tec}(x_t) + \sum_{c \in C} V_{teec}(x_t). \tag{1}$$

Fig. 2 shows a diagrammatic representation of a MRF modeling. Fig. 2(a) shows that each site s at location (i,j) is a MRF modeled with its neighbors in spatial direction. Fig. 2(b) shows the diagram for another MRF model in temporal direction. Here each site s at location (i,j) in the tth frame is modeled with neighbors of the corresponding pixels in the temporal direction, i.e., in the $(t-1)$th and $(t-2)$th frames. Similarly a MRF model that takes care of edge features is considered by modeling the line field of the tth frame with the neighbors of the corresponding pixels in the $(t-1)$th and $(t-2)$th frames. The MRF model diagram for line field is provided in Fig. 2(c).

B. MAP Estimation Based Framework for Initial Frame Segmentation

The observed image sequence y is assumed to be a degraded version of the actual image sequence x. For example at a given time t, the observed frame y_t is considered as a degraded version of the true label x_t. The degradation process is assumed to be Gaussian. Thus, the label field X_t can be estimated from the observed random field Y_t. The label field is estimated by maximizing the following posterior probability distribution:

$$\begin{aligned} \hat{x}_t &= \arg\ \max_{x_t}\ P(X_t = x_t | Y_t = y_t) \\ &= \arg\ \max_{x_t}\ \frac{P(Y_t=y_t|X_t=x_t)P(X_t=x_t)}{P(Y_t=y_t)} \end{aligned} \tag{2}$$

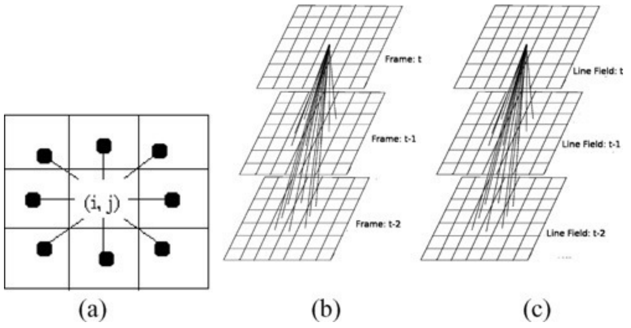

Fig. 2. (a) Neighborhood of a site for MRF modeling in the spatial direction. (b) MRF modeling taking two previous frames in the temporal direction. (c) MRF with two additional frames with line fields to take care of edge features.

where $\hat{x}_t$ denotes the estimated labels. The prior probability $P(Y_t = y_t)$ is constant and hence (2) reduces to

$$\hat{x}_t = \arg \max_{x_t} \; P(Y_t = y_t | X_t = x_t, \theta) P(X_t = x_t, \theta) \qquad (3)$$

where θ is the parameter vector associated with the clique potential function of x_t. The prior probability $P(X_t = x_t)$ can be expressed as

$$P(X_t = x_t) = e^{-U(x_t)}$$
$$= e^{-\left\{\sum_{c\in C} V_{sc}(x_t)+\sum_{c\in C} V_{tec}(x_t)+\sum_{c\in C} V_{teec}(x_t)\right\}}. \qquad (4)$$

In (4), $V_{sc}(x_t)$ is the clique potential function in spatial domain at time t, $V_{tec}(x_t)$ denotes the clique potential in temporal domain, and $V_{teec}(x_t)$ denotes the clique potential in temporal domain with edge features.

Assuming decorrelation of the three RGB planes for the color image and the variance to be the same among each plane, the likelihood function $P(Y_t = y_t | X_t = x_t)$ can be expressed as

$$P(N = y_t - x_t | X_t, \theta) = \frac{1}{\sqrt{(2\pi)^3 \sigma^3}} e^{-\frac{1}{2\sigma^2}(y_t - x_t)^2}. \qquad (5)$$

In (5), variance σ^2 corresponds to the Gaussian degradation. Here n is a realization of the Gaussian noise $N(\mu, \sigma^2)$.

Using (4), (5) and the underlying assumption of the degradation process, (3) reduces to

$$\hat{x}_t = \arg \min_{x_t} \left[\frac{\| y_t - x_t \|^2}{2\sigma^2} \right]$$
$$+ \left[\sum_{c\epsilon C} V_{sc}(x_t) + V_{tec}(x_t) + V_{teec}(x_t) \right]. \qquad (6)$$

$\hat{x}_t$ is the MAP estimate. The complete derivation is provided in Appendix A.

1) *Hybrid Algorithm for MAP Estimation of Initial Frame:* There are two kinds of relaxation schemes. This includes relaxation labeling and probabilistic relaxation. Relaxation labeling is very popular in MRF-MAP estimation and popularly referred to as stochastic relaxation [9]. In this regards SA, a generic probabilistic meta-heuristic optimization scheme proposed by Kirkpatrick *et al.* [22], is found to have a good approximation of the global optimum of a given function. Such an optimization scheme is inspired from the concept of annealing in metallurgy, a technique involving simultaneous heating and controlled cooling of a material, so that the particles of the material arrange themselves in the lower ground states of the corresponding lattice. In each step of SA, the algorithm replaces the current solution by a random "nearby" solution, chosen with a probability that depends on the difference between the corresponding functional values and the global parameter T (called the temperature), that is gradually decreased during the process. Initially, T is set to a high value, and in each step of processing it is decreased in a controlled manner. Hence, the computational time taken by SA is expected to be high. SA assumes that the cooling rate is low enough for the probability distribution of the current state to be near thermodynamic equilibrium at all times. This probability can be given as

$$P_r(U = u) = \frac{1}{Z} exp(-\frac{U}{k_B T}) \qquad (7)$$

where Z is the partition function, k_B is the Boltzman constant and U is the functional values or energy value. As reported by Li [18], the SA algorithm is a meta-heuristic search scheme, although optimize through the neighborhood searching approach, it can even move through the neighbors that are worse than the current solutions. SA is thus expected not to stop at a local optimum. In theory, if SA can run for an infinite amount of time, the global optimum could be found. It is also reported [18] that for any given finite problem, the SA terminates with the global optimal solution as the annealing schedule is extended.

ICM [23] uses a deterministic strategy to find the local minimum. It starts with an estimate of the labeling, and for each pixel, the label that gives a decrease in energy value is chosen for next iteration of processing. This process is repeated until convergence, which is guaranteed to occur, and in practice is very rapid. However, the results are extremely sensitive to the initial estimate, as it may stuck at local minima. The ICM algorithm is a deterministic scheme and starts with an initial label. In ICM algorithm for each pixel it searches for a neighborhood point that gives a decrease in energy function. Hence ICM may be viewed as a local search algorithm and the result provided by ICM algorithm may stuck to local minima. It is faster than SA.

We have proposed a hybrid algorithm (hybridization of both SA and ICM algorithms) for MAP estimation of the initial frame. The proposed algorithm works as follows: initially a few iterations of SA algorithm is executed to achieve a near optimal solution. Thereafter, for quick convergence, a local convergence based strategy, ICM, is run to converge to the nearest optimal solution.

The steps of the proposed algorithm are enumerated as below.

1) Initialize the temperature $T(t) = T_0$.
2) Compute the energy U of the configuration.
3) Perturb the system slightly with suitable Gaussian disturbance.
4) Compute the new energy $U^{'}$ of the perturbed system and evaluate the change in energy $\Delta U = U^{'} - U$.

5) If ($\Delta U < 0$), accept the perturbed system as a new configuration, else accept the perturbed system as a new configuration with probability $e^{-(\frac{\Delta U}{T(t)})}$.
6) Decrease the temperature $T(t+1) = c * T(t)$, where c is the cooling constant ($0 < c < 1$).
7) Repeat Steps 2–7 for some prespecified number of epochs.
8) Compute the energy U of the configuration.
9) Perturb the system slightly with suitable Gaussian disturbance.
10) Compute the new energy $U^{'}$ of the perturbed system and evaluate the change in energy $\Delta U = U^{'} - U$.
11) If ($\Delta U < 0$), accept the perturbed system as a new configuration, otherwise retain the original configuration.
12) Repeat Steps 8–12, till the stopping criterion $\Delta U < \epsilon$ (a predefined positive constant) is met.

C. Change Information Based Segmentation Scheme for Subsequent Frames

Spatio-temporal spatial segmentation in MRF-MAP framework is computation intensive due to random initialization. In order to reduce this burden, we propose a change information based heuristic initialization technique. This requires a previously segmented frame which is used in combination with the change information between the current and the previously considered frames to generate an initialization for processing the current frame. The change information is obtained by computing absolute values of the intensity difference between the current and the previously considered frames followed by a thresholding approach. The pixel values of the changed region of the current frame are superimposed on the previously available segmented frame to get an initialization.

Let y_t denote a frame at time t, whose spatio-temporal spatial segmentation x_t is available with us. Now considering y_{t+d} as a frame at an instant $(t+d)$, $x_{(t+d)i}$ represents its initialization obtained by this scheme. x_{t+d} represents its final spatio-temporal spatial segmentation. $x_{(t+d)i}$ can be obtained as follows.

1) Obtain the changed region corresponding to the frame y_{t+d} by taking a difference of the gray values of the frames y_{t+d} and y_t followed by thresholding, and the changes thus obtained is denoted by $y_{(t+d)_{|y_{t+d}-y_t|}}$.
2) The pixels corresponding to these changed regions in the tth frame segmentation result x_t are initialized as

$$x_{tt} = x_t - x_{t_{|y_{t+d}-y_t|}}. \tag{8}$$

These regions in the x_{tt} are replaced by the original gray values of y_{t+d} for initialization of the $(t+d)$th frame as

$$x_{(t+d)i} = x_{tt} + y_{(t+d)_{|y_{t+d}-y_t|}} \tag{9}$$

$y_{(t+d)_{|y_{t+d}-y_t|}}$ represents the pixels of the $(t+d)$th frame where changes took place from the previous frame. $x_{(t+d)i}$ serves as the initialization for spatio-temporal spatial segmentation of the $(t+d)$th frame. ICM is run on the $(t+d)$th frame starting from $x_{(t+d)i}$ to obtain $x_{(t+d)}$.

To illustrate the proposed technique, let us consider an example of *Bird* video. The original 27th and 31st frames are shown in Fig. 3(a) and (b). Segmentation result for the 27th frame, x_{27} using *edgebased* compound MRF model and hybrid algorithm is displayed in Fig. 3(c). By taking the absolute value of pixel by pixel intensity difference of 27th and 31st frames we obtain the difference image as shown in Fig. 3(d). The corresponding thresholded image is shown in Fig. 3(e). The pixel values of the 31st frame of the changed region is superimposed on the 27th segmented frame, i.e., x_{27} to generate x_{31i} [shown in Fig. 3(f)]. ICM is run on the 31st frame starting from x_{31i} to obtain the segmentation result for the 31st frame, i.e., x_{31} [as shown in Fig. 3(g)].

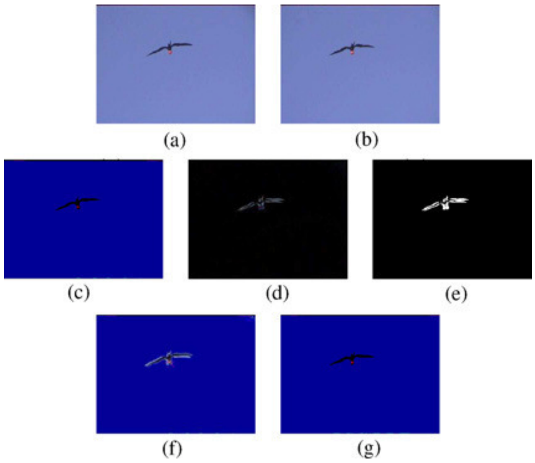

Fig. 3. *Bird* video. (a) Original frame 27. (b) Original frame 31. (c) *Edgebased* segmentation result of 27th frame. (d) Difference image obtained by pixel by pixel comparison of 27th and 31st frames. (e) Thresholded difference image for 31st frame. (f) Initialization for 31st frame. (g) Final segmentation result of 31st frame.

For MRF based segmentation, the set of all possible image configurations is given by $D = 2^{b \times M \times N}$, where b is the number of bits used to represent the gray value of each pixel in the image frame and $M \times N$ is the dimension of the image frame. 2^b represents the admissible pixel values and D represents all admissible realization of images in the $M \times N$ dimensional real space. The x-axis of the plot in Fig. 4 represents the number of possible image frames, that ranges from 0 to D. x_t represents one such realized image frame form this range. Searching such a huge space every time for each frame requires high computational time. (Note that the contents of the scene is not changing much from one frame to another.) To minimize this burden, we have considered the above approach where an initialization $x_{(t+d)i}$ is expected to lie near the optimum point e as shown in Fig. 4. Now considering a local (optimum) fast searching criterion, such as ICM, we can detect the optimum point o.

IV. Temporal Segmentation

Generally temporal segmentation is performed to classify the foreground and the background in a video image frame. In temporal segmentation, a CDM is obtained and this CDM serves as a precursor for detection of the foreground as well as the background. The general procedure for obtaining the CDM is by taking a difference between the gray value of the current and the previously considered frame followed by a thresholding algorithm. In order to detect moving objects

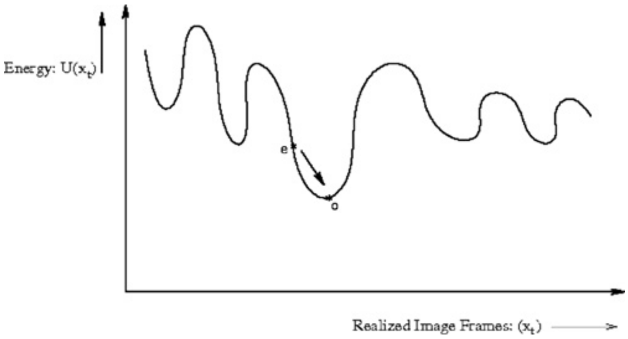

Fig. 4. Energy curve for each realized image through MRF.

in the absence of reference frame, some information from the previous frame is used to update the current CDM. As opposed to the conventional gray value difference CDM, we have considered a 'label frame' difference CDM, where silhouette effect is less. The label frame difference CDM is obtained by taking a difference of x_t and x_{t+d}.

The results thus obtained by a CDM with difference in label values of two frames are compared with those of the CDM constructed with a gray value difference of two frames. We have adopted a popular global thresholding method such as Otsu's method [30] for thresholding the CDM. The results thus obtained are verified and compensated by considering the information of the pixels belonging to the objects in the previous frame, to improve the segmentation result of the moving objects. This is represented as

$$R = \left\{ r_{i,j} | 0 \leq i \leq (M-1), 0 \leq j \leq (N-1) \right\} \tag{10}$$

where R is a matrix having the same size of the frame. $r_{i,j}$ is the value of the VOP at location (i, j), where (i, j) location represents the ith row and the jth column (detail explanation is available in Appendix B). If a pixel is found to have $r_{i,j} = 1$, then it belongs to a moving object in the previous frame; otherwise it belongs to the background in the previous frame. Based on this information, CDM is modified as follows: if it belongs to a moving object in the previous frame and its label obtained by spatio-temporal spatial segmentation is the same as that of one of the corresponding pixels in the previous frame, then it is marked as a foreground area in the current frame else as the background. The modified CDM thus represents the final temporal segmentation.

V. VOP Generation and Tracking

After obtaining a temporal segmentation of a frame at time t, we get a binary output with objects as one class (denoted by FM_t) and the background as other class (denoted as BM_t).

The regions forming the foreground part in the temporal segmentation is identified as moving object regions, and the pixels corresponding to the FM_t part of the original frame y_t form the VOP. After obtaining the VOP from the different image frames we can track the moving objects.

After obtaining a temporal segmentation, a centroid based tracking is performed to track moving objects from the considered video image sequence. The centroid $(\hat{u}_{n_c}, \hat{v}_{n_c})$ of the moving object is computed as

$$\hat{u}_{n_c} = \frac{\sum_i u_{n_i} c(i)}{\sum_{i \epsilon T} c(i)} \tag{11}$$

$$\hat{v}_{n_c} = \frac{\sum_i v_{n_i} c(i)}{\sum_{i \epsilon T} c(i)} \tag{12}$$

where (u_{ni}, v_{ni}) represents the co-ordinate of a pixel in the temporal segmentation and the value of $c(i)$ is considered as

$$c(i) = \begin{cases} 1, & \text{if pixel i is identified as an object pixel} \\ 0, & \text{if pixel i is identified as a background pixel.} \end{cases}$$

By calculating the $(\hat{u}_{n_c}, \hat{v}_{n_c})$ for different frames, the movement of an object can be tracked.

VI. Experimental Result and Analysis

We have considered two types of video sequences (one reference video sequence and one real life video sequence) as shown in Figs. 5 and 6 to test the usefulness of the proposed approach. Since changes in between the consecutive frames are very less, we have considered a few randomly sampled frames within a particular interval of time where a reasonable amount of change is expected to have occurred. For the given video sequences the spatial segmentation of the initial frame has been obtained by the proposed *edgebased* compound MRF model followed by the hybrid algorithm for MAP estimation. For the subsequent frames the spatial segmentation is obtained using the change information based initialization scheme. The label frame difference based temporal segmentation and the spatio-temporal spatial segmentation is combined to obtain the VOP. Thereafter tracking is carried out. In order to validate the scheme, we have compared the spatio-temporal spatial segmentation results obtained by the proposed scheme with those of JSEG [24], *edgeless* [25] and *edgebased* [25] methods of segmentation and is found to be better in terms of numbers of misclassified pixels. The computational time requirement for the proposed method is also found to be less as compared to *edgeless* and *edgebased* approaches of segmentation. Similarly the results obtained for VOP by the label frame difference CDM is compared with those of CDM with a gray level difference, and it is found that the VOP with label frame difference CDM approach gives better results opposed to gray level difference CDM.

The first video considered is the *Akiyo* video sequence. Fig. 5(a) shows the original image frames of this video sequence. Corresponding manually constructed ground truth images are shown in Fig. 5(b). The initial frame of this video is segmented by modeling it with the proposed *edgebased* compound MRF model followed by the hybrid algorithm for MAP estimation. The proposed change information based subsequent frame segmentation scheme is used to segment the different frames of this sequence. Fig. 5(c) shows the spatial segmentation of these image frames using the proposed change information scheme. The MRF model parameters chosen for this video are $\alpha = 0.009$, $\beta = 0.008$, $\gamma = 0.007$ and $\sigma = 2.0$. Segmentation results of these frames using *edgeless* approach are shown in Fig. 5(d). It is observed from these figures that the racks behind *Akiyo* are not properly segmented. The distinction in the shapes of the racks are not proper. Similarly, the face of *Akiyo* is not properly segmented (e.g., lip, eye). The blazer crest and the shirt like portion are almost lost. Segmentation result of JSEG approach are shown in Fig. 5(e), which also fails to segment the frames properly and an over-segmented result in the portions such as face, racks, blazer, shirt and others are obtained. Except the display portion all portions in the background are merged into a single class. The proposed approach however is able to segment the face, lip, eye, and others properly. Similarly, the distinction in the shapes of the racks with minute edge details are properly identified by the proposed change information based segmentation scheme.

The temporal segmentation results of these frames, obtained using the CDM generated with a difference in label frames instead of the CDM generated with a difference in original frames are shown in Fig. 5(f) and the corresponding VOPs are shown in Fig. 5(g). It is observed from these VOPs that the object (i.e., *Akiyo*) in different frames has been detected properly. The corresponding temporal segmentation results using a difference in original frame based CDM are shown in Fig. 5(i). It is observed from these results that there are some patches near the hair portions of *Akiyo* which led to the misclassification in VOPs [shown in Fig. 5(j)]. Thus, we notice that temporal segmentation obtained using the CDM generated with a label frame difference yields better VOPs than that of using a gray value difference CDM. Result of tracking is shown in Fig. 5(h).

The segmentation result using MRF modeling with *edgeless* approach is shown in Fig. 5(d). The results obtained by JSEG [24] method are shown in Fig. 5(e). The number of misclassified pixels is computed by comparing the result thus obtained with the ground truth, and is shown in Table I. It can be observed that the proposed method incurs less misclassification than that of using JSEG method and the *edgeless* approach of segmentation. But the results are quite comparable with *edgebased* approach of spatial segmentation.

In order to test the robustness of the proposed algorithm, we also tested it on one real life video sequence i.e., (Fig. 6) with uncontrolled environmental conditions. Fig. 6 represents the VOP generated for *Rahul* video sequence. This video was captured with a low resolution video camera at the National Institute of Technology, Rourkela, India. Fig. 6(a) represents the original image frames of this video sequence. Corresponding ground truth image frames are shown in Fig. 6(b). Fig. 6(c) shows the spatio-temporal spatial segmentation result of these frames by the proposed spatial segmentation scheme. The MRF model parameters chosen for this video are α = 0.009, β = 0.005, γ = 0.001 and σ = 5.0. Spatial segmentation results of those frames using *edgeless* and JSEG approaches of segmentation are shown in Fig. 6(d) and (e), respectively. It is observed from the results that these methods provide over segmented results in the face and hand regions of *Rahul*. The JSEG scheme segments the lower parts of the hand of *Rahul* in the background class. Similarly, some portions like collar of *Rahul* is merged with face regions of *Rahul*. Fig. 6(g) shows the generated VOPs of *Rahul* using label frame difference CDM. The results for VOPs of *Rahul* video sequence using gray level difference based CDM are shown in Fig. 6(j). It is observed that the effect of silhouette is quite less in Fig. 6(g) than that in Fig. 6(j). Result of tracking is shown in Fig. 6(h).

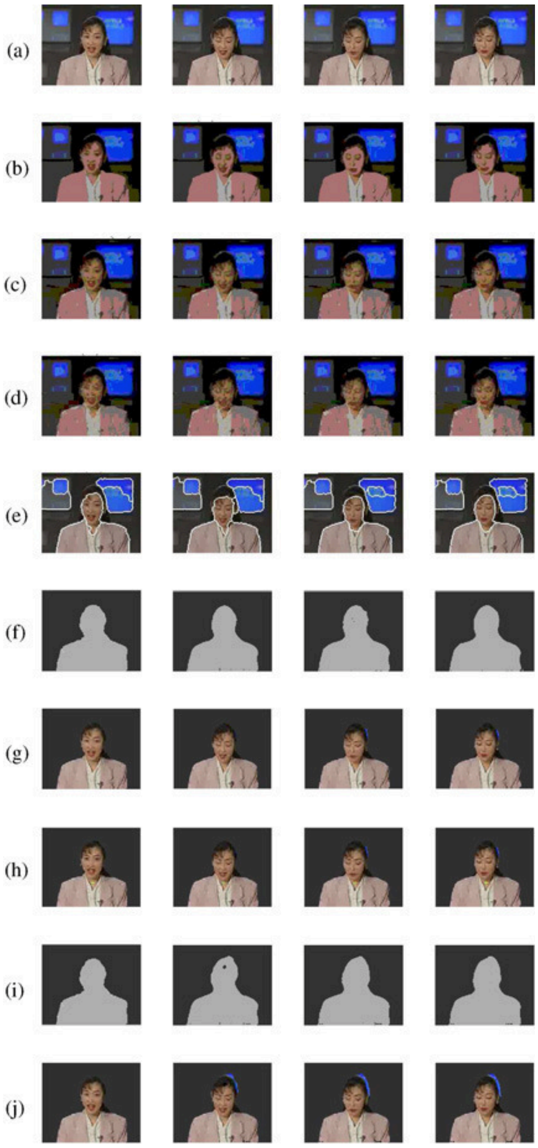

Fig. 5. VOP generation for *Akiyo* video sequence using change information based scheme (for frames 75th, 95th, 115th, and 135th). (a) Original frames. (b) Ground truth of original frames. (c) Segmentation using proposed scheme. (d) Segmentation using edgeless scheme. (e) Segmentation using JSEG scheme. (f) Temporal segmentation using label frame CDM. (g) VOP generated by temporal segmentation result (f). (h) Centroid based tracking of VOPs as obtained in (g). (i) Temporal segmentation using original frame CDM. (j) VOP generated by temporal segmentation result (i).

We have used a single set of SA parameters for all the video sequence. Those are initial temperature (T_0) = 0.38, cooling constant (c) = 0.9992, perturbation variance as 1.

A. Computational Time Requirement

All the programs were implemented in a Pentium 4(D), 3 GHz, L2 cache 4 MB, 1 GB RAM, 667 FSB PC with Fedora-Core operating system and C programming language. The image sequence considered are of size 176 × 144.

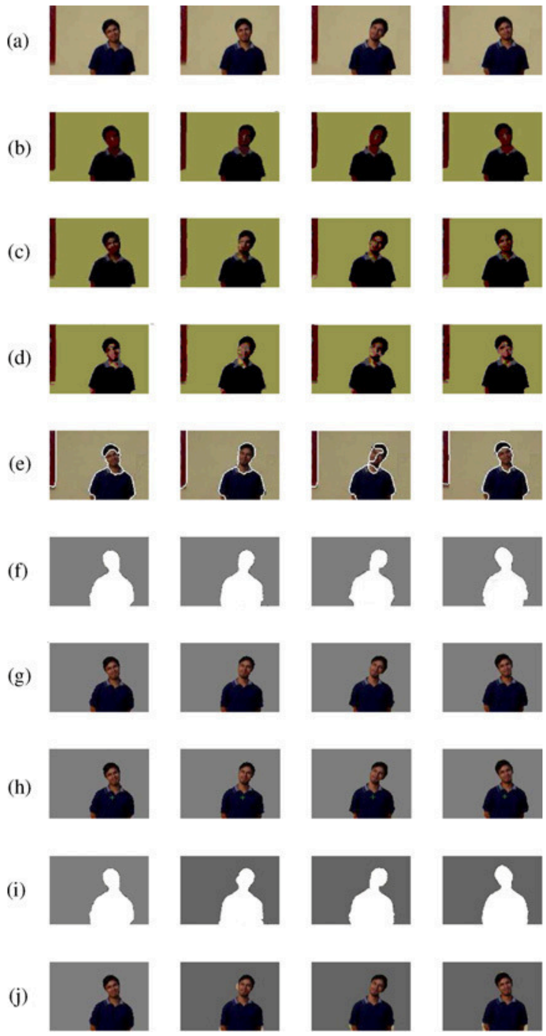

Fig. 6. VOP generation for *Rahul* video using change information based scheme (for frames 11th, 16th, 21st, and 26th). (a) Original frames. (b) Ground truth of original frames. (c) Segmentation using proposed scheme. (d) Segmentation using edgeless scheme. (e) Segmentation using JSEG scheme. (f) Temporal segmentation using label frame CDM. (g) VOP generated by temporal segmentation result (f). (h) Centroid based tracking of VOPs as obtained in (g). (i) Temporal segmentation using original frame CDM. (j) VOP generated by temporal segmentation result (i).

The time required by the proposed scheme, to detect the moving objects from the considered video sequences, are provided in Table II. We have considered a few sample frames of a particular video sequence and tested our algorithm on it and the average time taken is assumed as $t1$. We tested *edgebased* and *edgeless* approaches of segmentation on the same set of image frames and the average time taken by each of them is denoted as $t2$; and have computed the *Gain* by using the following formula:

$$Gain = \frac{t2}{t1}.$$

We have computed the *Gain* for the two video examples. For *Akiyo* sequence it comes 10.25 with *edgebased* and 13.4 with *edgeless* approach. Similarly, for *Rahul* video sequence it is 8.5 with *edgebased* and 10 with *edgeless* approach.

TABLE I
NUMBER OF MISCLASSIFIED PIXELS

Video	*FrameNo.*	*Edgeless*	*Edgebased*	*Proposed*	*JSEG*
Akiyo	75	388	88	88	4718
	95	312	75	75	1238
	115	259	106	115	1262
	135	335	91	115	1374
Rahul	11	100	51	51	300
	15	115	93	85	310
	16	102	68	72	380
	21	112	63	66	308

TABLE II
TIME (IN SECOND) REQUIRED FOR EXECUTION OF THE ALGORITHMS PER FRAME

Video	*Frame No.*	*Edgeless*	*Edgebased*	*Proposed*
Akiyo	95	108	82	8
	115	108	82	8
	135	108	82	8
Rahul	16	70	57	7
	21	70	57	7
	26	70	57	7

From the considered video examples we observed that using change information based spatio-temporal approach a better accuracy of segmentation is obtained with a faster execution time. Similarly, using a label frame difference CDM instead of gray level difference CDM effect of silhouette is found to be reduced. Thus the change information based scheme has much less computational burden and gives more accuracy, and hence is more viable for real time implementation. Since JSEG approach does yield an over-segmented result, which is unacceptable, we have not compared the execution time of JSEG scheme with the proposed approach.

VII. CONCLUSION AND DISCUSSION

In this article, a change information based moving object detection scheme is proposed. The spatio-temporal spatial segmentation result of the initial frame is obtained by *edgebased* MRF modeling and a hybrid MAP estimation algorithm (hybrid of SA and ICM). The segmentation result of the initial frame together with some change information from other frames is used to generate an initialization for segmentation of other frames. Then, an ICM algorithm is used on that frame starting from the obtained initialization for segmentation. It is found that the proposed approach produces better segmentation results compared to those of *edgeless* and JSEG segmentation schemes and comparable results with *edgebased* approach. The proposed scheme gives better accuracy and is approximately 13 times faster compared to the considered MRF based segmentation schemes for a number of video sequences. The MRF model parameters are chosen on a trial and error basis. For temporal segmentation a CDM based on a difference of labels of two frames is considered. This reduces the effect of silhouette on the generated VOP. A centroid based tracking process is considered to track the objects.

Our future work will focus on estimation of MRF model parameters and VOP generation of the initial frame. We are also looking at a related problems with moving camera where existing approach does not produce good results.

Appendices

A. Appendix

Here image segmentation problem is considered to be a process of determining a realization x_t that has given rise to the actual image frame y_t. The realization x_t cannot be obtained deterministically from y_t. Hence, it requires to estimate $\hat{x}_t$ from y_t. One way to estimate $\hat{x}_t$ is based on the statistical MAP criterion. The objective of statistical MAP estimation scheme is to have a rule, which yields $\hat{x}_t$ that maximizes the a posteriori probability, that is

$$\hat{x}_t = \arg\max_{x_t} P(X_t = x_t | Y_t = y_t) \tag{13}$$

where $\hat{x}_t$ denotes the estimated labels. Since x_t is unknown, it is difficult to evaluate (13). Using Bayes' theorem, (13) can be written as

$$\hat{x}_t = \arg\max_{x_t} \frac{P(Y_t = y_t | X_t = x_t) P(X_t = x_t)}{P(Y_t = y_t)}. \tag{14}$$

Since y_t is known, the prior probability $P(Y_t = y_t)$ is constant. Hence (14) reduces to

$$\hat{x}_t = \arg\max_{x_t} P(Y_t = y_t | X_t = x_t, \theta) P(X_t = x_t, \theta) \tag{15}$$

where θ is the parameter vector associated with the clique potential function of x_t. According to Hammerseley Clifford theorem [18], the prior probability $P(X_t = x_t, \theta)$ follows Gibb's distribution and is of the following form:

$$\begin{aligned} P(X = x) &= e^{-U(x,\theta)} \\ &= e^{\left[-\sum_{c\epsilon C}[V_{sc}(x)+V_{tec}(x)+V_{teec}(x)]\right]}. \end{aligned} \tag{16}$$

In (16), $V_{sc}(x_t)$ is the clique potential function in spatial domain at time t, $V_{tec}(x_t)$ denotes the clique potential in temporal domain, and $V_{teec}(x_t)$ denotes the clique potential in temporal domain with edge features. We have used this additional feature in the temporal direction and the whole model is referred to as *edgebased* model. The corresponding *edgeless* model is expressed as

$$P(X_t = x_t) = e^{-U(x_t,\theta)} = e^{\left[-\sum_{c\epsilon C}[V_{sc}(x_t)+V_{tec}(x_t)]\right]}. \tag{17}$$

The likelihood function $P(Y_t = y_t | X_t = x_t)$ of (15) can be expressed as

$$\begin{aligned} P(Y_t = y_t | X_t = x_t) &= P(y_t = x_t + n | X_t = \\ & x_t, \theta) = P(N = y_t - x_t | X_t = x_t, \theta). \end{aligned}$$

Here n is a realization of the Gaussian degradation process $N(\mu, \sigma)$. Thus, $P(Y_t = y_t | X_t = x_t)$ can be expressed as

$$\begin{aligned} & P(N = y_t - x_t | X_t, \theta) \\ &= \frac{1}{\sqrt{(2\pi)^f det\,[k]}} e^{-\frac{1}{2}(y_t - x_t)^T k^{-1} (y_t - x_t)} \end{aligned} \tag{18}$$

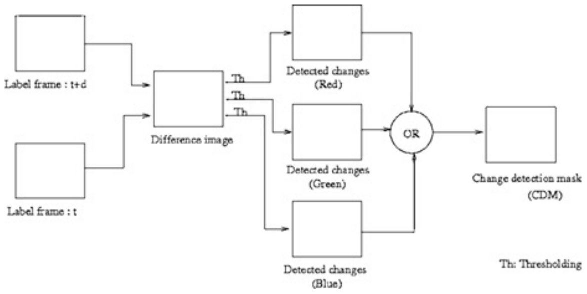

Fig. 7. Blockdiagram of temporal segmentation scheme.

where k is the covariance matrix, $det[k]$ represents the determinant of matrix k and f is the number of features (for color image, RGB are the three features). Assuming decorrelation among the three RGB planes and the variance to be the same among all the planes, (6) can be expressed as

$$P(N = y_t - x_t | X_t, \theta) = \frac{1}{\sqrt{(2\pi)^3 \sigma^3}} e^{-\frac{1}{2\sigma^2}(y_t - x_t)^2}. \tag{19}$$

In (19), variance σ^2 corresponds to the Gaussian degradation. Hence (15) can be expressed as

$$\begin{aligned} \hat{x}_t = \arg\max_{x_t} & \frac{1}{\sqrt{(2\pi)^3 \sigma^3}} \times \\ & \left[e^{\frac{\left[-\| y_t - x_t \|^2\right]}{2\sigma^2} - \left[\sum_{c\epsilon C}[V_{sc}(x_t)+V_{tec}(x_t)+V_{teec}(x_t)]\right]} \right]. \end{aligned} \tag{20}$$

Maximization of (20) is equivalent to minimization of

$$\begin{aligned} \hat{x}_t = \arg\min_{x_t} & \left[\frac{\| y_t - x_t \|^2}{2\sigma^2} \right] + \\ & \left[\sum_{c\epsilon C} V_{sc}(x_t) + V_{tec}(x_t) + V_{teec}(x_t) \right]. \end{aligned} \tag{21}$$

$\hat{x}_t$ in (21) is the MAP estimate.

B. Appendix

For temporal segmentation, we have obtained the label difference image by taking a difference of the respective R, G and B components of the considered frames' spatial segmentation result. We applied Otsu's thresholding algorithm [30] to each channel (i.e., R, G and B) of the difference image. After obtaining the thresholded images for all the channels, they were fused by a logical operator (we have considered the OR operator here). The schematic representation of the above process is shown in Fig. 7.

The CDM thus obtained are of two types, either changed or unchanged (i.e., denoted as 1 or 0). To improve the temporal segmentation, the obtained change detection output is combined with the VOP of the previous frame based on the label information of the current and previous frames. The VOP of the previous frame is represented as a matrix of size $M \times N$ as

$$R = \left\{ r_{i,j} | 0 \leq i \leq (M-1), 0 \leq j \leq (N-1) \right\} \tag{22}$$

and is represented as

$$R = \begin{pmatrix} r_{0,0} & r_{0,1} & r_{0,2} & \cdot & \cdot & \cdot & r_{0,N-1} \\ r_{1,0} & r_{1,1} & r_{1,2} & \cdot & \cdot & \cdot & r_{1,N-1} \\ r_{2,0} & r_{2,1} & r_{2,2} & \cdot & \cdot & \cdot & r_{2,N-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{M-1,0} & r_{M-1,1} & r_{M-1,2} & \cdot & \cdot & \cdot & r_{M-1,N-1} \end{pmatrix}$$

where $r_{i,j}$ is the value of the VOP at location (i, j). Here (i, j) location represents the ith row and the jth column and is described as

$$r_{i,j} = \begin{cases} 1, & \text{if it is in object} \\ 0, & \text{if it is in background.} \end{cases} \tag{23}$$

Acknowledgments

The authors would like to acknowledge the thorough and constructive comments provided by the reviewers and the editors in this paper. They would also like to thank the Department of Science and Technology, Government of India, and University of Trento, Trento, Italy, the sponsors of the ITPAR Program.

References

[1] A. L. Bovic, *Image and Video Processing*. New York: Academic Press, 2000.

[2] A. M. Tekalp, *Digital Video Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[3] Y. Amit, *2-D Object Detection and Recognition*. Cambridge, MA: MIT Press, 2002.

[4] R. F. Gonzalez and R. E. Wood, *Digital Image Processing*. Singapore: Pearson Education, 2001.

[5] P. Salember and F. Marques, "Region based representation of image and video segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1147–1169, Dec. 1999.

[6] A. Blake and A. Zisserman, *Visual Reconstruction*. London, U.K.: MIT Press, 1987.

[7] K. S. Shanmugam and A. M. Breipohl, *Random Signals Detection, Estimation and Data Analysis*. New York: Wiley, 1988.

[8] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Process*. New York: McGraw-Hill, 2002.

[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[10] A. Ghosh, N. R. Pal, and S. K. Pal, "Image segmentation using a neural network," *Biol. Cybern.*, vol. 66, no. 2, pp. 151–158, 1991.

[11] R. O. Hinds and T. N. Pappas, "An adaptive clustering algorithm for segmentation of video sequences," in *Proc. ICASSP*, vol. 4. 1995, pp. 2427–2430.

[12] G. K. Wu and T. R. Reed, "Image sequence processing using spatiotemporal segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 5, pp. 798–807, Aug. 1999.

[13] E. Y. Kim, S. W. Hwang, S. H. Park, and H. J. Kim, "Spatiotemporal segmentation using genetic algorithms," *Pattern Recognition*, vol. 34, no. 10, pp. 2063–2066, 2001.

[14] S. W. Hwang, E. Y. Kim, S. H. Park, and H. J. Kim, "Object extraction and tracking using genetic algorithms," in *Proc. Int. Conf. Image Process.*, vol. 2. 2001, pp. 383–386.

[15] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: A region labeling approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 7, pp. 597–612, Jul. 2002.

[16] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Process.*, vol. 11, no. 5, pp. 530–544, May 2002.

[17] E. Y. Kim and S. H. Park, "Automatic video segmentation using genetic algorithms," *Pattern Recognition Lett.*, vol. 27, no. 11, pp. 1252–1265, 2006.

[18] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. New York: Springer, 2001.

[19] S. D. Babacan and T. N. Pappas, "Spatiotemporal algorithm for joint video segmentation and foreground detection," in *Proc. EUSIPCO*, Sep. 2006, pp. 1–5.

[20] S. D. Babacan and T. N. Pappas, "Spatiotemporal algorithm for background subtraction," in *Proc. IEEE ICASSP*, Apr. 2007, pp. 1065–1068.

[21] S. S. Huang and L. Fu, "Region-level motion-based background modeling and subtraction using MRFs," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1446–1456, May 2007.

[22] S. C. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[23] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986.

[24] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[25] B. N. Subudhi and P. K. Nanda, "Compound Markov random field model based video segmentation," in *Proc. SPIT-IEEE Colloq. Int. Conf. 2007–2008*, vol. 1. Feb. 2008, pp. 97–102.

[26] E. Durucan and T. Ebrahimi, "Moving object detection between multiple and color images," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, Jul. 2003, pp. 243–251.

[27] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.

[28] N. Anjum and A. Cavallaro, "Multifeature object trajectory clustering for video analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1555–1564, Nov. 2008.

[29] B. G. Kim, D. J. Kim, and D. J. Park, "Novel precision target detection with adaptive thresholding for dynamic image segmentation," *Mach. Vis. Applicat.*, vol. 12, no. 5, pp. 259–270, 2001.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[31] G. H. Sasaki and B. Hajek, "The time complexity of maximum matching by simulated annealing," *J. ACM*, vol. 35, no. 2, pp. 387–403, 1988.

Badri Narayan Subudhi received the B.E. degree in electronics and telecommunication from the Biju Patnaik University of Technology, Rourkela, India, in 2004, and the M.Tech. degree in electronics system and communication from the National Institute of Technology, Rourkela, in 2009.

He is currently a Research Scholar with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His current research interests include video processing, image processing, machine learning, pattern recognition, and remote sensing image analysis.

Pradipta Kumar Nanda (M'08) received the B.Tech. degree from the VSS University of Technology, Sambalpur, India, in 1984, the Masters degree in electronics and communication engineering from the National Institute of Technology, Rourkela (NIT Rourkela), India, in 1989, and the Ph.D. degree in the area of computer vision from the Indian Institute of Technology Bombay, Mumbai, India, in 1996.

He was a Professor with NIT Rourkela and has served the institute for 21 years before moving to Siksha O Anusandhan University, Bhubaneswar, India. He is currently a Professor with the Department of Electronics and Telecommunication Engineering, and is the Chairman of the Research and Development Cell of the Institute of Technical Education and Research, Siksha O Anusandhan University. His current research interests include image processing and analysis, bio-medical image analysis, video-tracking, soft computing, and its applications.

Dr. Nanda is a fellow of IETE, India.

Ashish Ghosh (M'09) received the B.E. degree in electronics and telecommunication from Jadavpur University, Kolkata, India, in 1987, and the M.Tech. and Ph.D. degrees in computer science from the Indian Statistical Institute, Kolkata, in 1989 and 1993, respectively.

He is currently a Professor with the Machine Intelligence Unit, Indian Statistical Institute. He was selected as an Associate of the Indian Academy of Sciences, Bangalore, India, in 1997. He visited Osaka Prefecture University, Osaka, Japan, with a Post-Doctoral Fellowship from October 1995 to March 1997, and Hannan University, Osaka, as a Visiting Faculty Member from September 1997 to October 1997 and from September 2004 to October 2004. He has visited Hannan University as a Visiting Professor with a fellowship from the Japan Society for Promotion of Sciences from February 2005 to April 2005. In May 1999, he was with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, with the CIMPA (France) Fellowship. He was with the German National Research Center for Information Technology, Germany, with a German Government Fellowship from January 2000 to April 2000, and with Aachen University, Aachen, Germany, in September 2010 with an European Commission Fellowship. From October 2003 to December 2003, he was a Visiting Professor with the University of California, Los Angeles, and from December 2006 to January 2007 he was with the Department of Computer Science, Yonsei University, Seoul, Korea. His visits to the University of Trento, Trento, Italy, and the University of Palermo, Palermo, Italy, from May 2004 to June 2004, March 2006 to April 2006, May 2007 to June 2007, 2008, 2009, and 2010, all in connection with collaborative international projects. He also visited various universities/academic institutes and delivered lectures in different countries, including Poland and the Netherlands. He has already published more than 120 research papers in internationally reputed journals and refereed conferences, and has edited eight books. His current research interests include pattern recognition and machine learning, data mining, image analysis, remotely sensed image analysis, video image analysis, soft computing, fuzzy sets and uncertainty analysis, neural networks, evolutionary computation, and bioinformatics.

Dr. Ghosh received the prestigious and most coveted Young Scientists Award in Engineering Sciences from the Indian National Science Academy in 1995, and in Computer Science from the Indian Science Congress Association in 1992. He is a member of the founding team that established the National Center for Soft Computing Research at the Indian Statistical Institute, Kolkata, in 2004, with funding from the Department of Science and Technology, Government of India, and is currently in charge of the Center. He is acting as a member of the editorial boards of various international journals.