Pattern Recognition Letters 32 (2011) 2097–2108

Entropy based region selection for moving object detection

Badri Narayan Subudhi [a], Pradipta Kumar Nanda [b], Ashish Ghosh [a,*]

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India
[b] Department of Electronics and Telecommunication Engineering, ITER, Siksha 'O' Anusandhan University, Bhubaneswar 751030, India

ARTICLE INFO

Article history:
Received 20 August 2010
Available online 22 August 2011
Communicated by S. Todorovic

Keywords:
Object detection
MAP estimation
Simulated annealing
Entropy
Thresholding
Gaussian distribution

ABSTRACT

This article addresses a problem of moving object detection by combining two kinds of segmentation schemes: temporal and spatial. It has been found that consideration of a global thresholding approach for temporal segmentation, where the threshold value is obtained by considering the histogram of the difference image corresponding to two frames, does not produce good result for moving object detection. This is due to the fact that the pixels in the lower end of the histogram are not identified as changed pixels (but they actually correspond to the changed regions). Hence there is an effect on object background classification. In this article, we propose a local histogram thresholding scheme to segment the difference image by dividing it into a number of small non-overlapping regions/windows and thresholding each window separately. The window/block size is determined by measuring the entropy content of it. The segmented regions from each window are combined to find the (entire) segmented image. This thresholded difference image is called the change detection mask (CDM) and represent the changed regions corresponding to the moving objects in the given image frame. The difference image is generated by considering the label information of the pixels from the spatially segmented output of two image frames. We have used a Markov Random Field (MRF) model for image modeling and the maximum a posteriori probability (MAP) estimation (for spatial segmentation) is done by a combination of simulated annealing (SA) and iterated conditional mode (ICM) algorithms. It has been observed that the entropy based adaptive window selection scheme yields better results for moving object detection with less effect on object background (mis) classification. The effectiveness of the proposed scheme is successfully tested over three video sequences.

1. Introduction

Moving objects detection from a given video is always a challenging task in video processing (Tekalp, 1995) and computer vision (Forsyth and Ponce, 2003). It has wide applications in diverse fields like: visual surveillance (Hu et al., 2004; Schiele et al., 2009), event detection (Chuang et al., 2009), activity recognition (Beleznai et al., 2006), activity based human recognition (Veeraraghavan et al., 2005), face and gait-based human recognition (Huang et al., 1999; Shi et al., 2008), fault diagnosis (Verma et al., 2004), path detection (Makris and Ellis, 2002), robotics (Satake and Miura, 2009), image and video indexing (Yong et al., 2007), etc. Based on the movements of objects and background, video sequences can be categorized into two types: moving objects with moving background, and moving objects with fixed background. In the later case moving object detection can be accomplished by motion detection (Tekalp, 1995) alone, where a reference frame is available. It uses frame subtraction (Gonzalez and Woods, 2001) or background subtraction scheme (Stauffer and Grimson, 1999) for detecting the changes between the background and the considered scenes. If the reference frame is not available and the objects in the scene have a significant amount of motion, then the background subtraction scheme may able to detect the objects in the scene. However, without availability of reference frame if the objects in the scene has a very slow movement or the objects in the scene stop for some time and move further, detection of moving objects from such a scene is very difficult. A combination of temporal segmentation (Tekalp, 1995) and spatial segmentation (Gonzalez and Woods, 2001) proved to be a better approach for such situations (Zhang, 2006).

Spatial segmentation may be defined as a process of dividing an image frame into a number of non-overlapping meaningful homogenous regions (Gonzalez and Woods, 2001). In the last two decades several spatial segmentation schemes have been developed in Computer Vision (Haralick and Shapiro, 1992) paradigm. Few popular schemes in this paradigm are thresholding based (Gonzalez and Woods, 2001), and region based (e.g. region growing (Zucker, 1976) and watershed (Kim et al., 1999)), clustering based (Comaniciu and Meer, 2002), MRF based (Li, 2001), etc. Thresholding based approaches do not exploit the spatio-contextual information of pixels and hence sometimes produce

* Corresponding author. Tel.: +91 33 2575 3110/3100; fax: +91 33 2578 3357.
E-mail addresses: subudhi.badri@gmail.com (B.N. Subudhi), pknanda_d13@ya hoo.co.in (P.K. Nanda), ash@isical.ac.in (A. Ghosh).

disjoint regions rather than a complete region. In contrast, region based schemes and MRF based schemes use spatio-contextual information and hence mostly yield acceptable segments. Often, noise and illumination variation in the scene alter the gray values of pixels and segmentation by spatio-contextual information based approaches produce over segmented results. Grey level information along with edge/contrast information proves to be a better approach towards this. A robust segmentation scheme using mean-shift clustering combined with edge information has also been used (Comaniciu and Meer, 2002).

In an image frame both stationary and non-stationary objects with background may be present, hence *spatial segmentation* scheme may provide the region boundary for both stationary and non-stationary objects. As our objective is to find the non-stationary objects only, a combination of both *spatial* and *temporal* cues of a given video may efficiently predict the locations of the moving objects. A moving object detection approach, that uses watershed based spatial segmentation coupled with motion detection is initially suggested by Kim et al. (1999). This method has two main drawbacks: it produces an over-segmented result and the complexity is very high. To enhance the accuracy of moving object detection, Deng and Manjunath (2001) proposed a robust scheme where spatial segmentation of an image frame is obtained by color quantization followed by region growing. This method is popularly known as joint segmentation (JSEG) scheme. For temporal segmentation, regions corresponding to objects are matched in the temporal direction by computing the motion vectors of the object regions in the target frame. It has been found that the use of region based segmentation scheme fails to take care of the spatial ambiguities of image gray values. Hence produces an over segmented result that gives rise to a large effect of object background misclassification for moving object detection. This is termed as "effects of silhouette" (Subudhi et al., 2009).

The gray level of pixels with high uncertainty and high ambiguity make it difficult to detect moving objects with more accuracy by non-statistical spatial segmentation methods. Hence, it requires some kind of stochastic method to model the important attributes of an image frame so that a better segmentation result can be obtained. Markov Random Field (MRF) model (Geman and Geman, 1984; Li, 2001), in this context, is proved to be a better framework. An early work on MRF based object detection scheme is proposed by Hinds and Pappas (1995) where temporal constraints and temporal local intensity adaptations are introduced to obtain a smooth transition of segmentation results from one frame to another. A robust spatio-temporal segmentation based moving object detection scheme is proposed by Hwang et al. (2001), where the spatial segmentation of each frame of the given video sequence is obtained by attribute modeling with MRF and MAP estimation by distributed genetic algorithm (DGA). Temporal segmentation is obtained by direct combination of video object plane (VOP) of the previous frame with the current frame's change detection mask (CDM). For further improvement in object detection and to reduce object and background misclassification, Kim and Park (2006) extended the work in (Hwang et al., 2001). To reduce the computational time of MRF–DGA, the authors have used results of previous frame segmentation as a cue for subsequent frame segmentation. A probabilistic frame-work (termed as evolutionary probability) is used here to update the crossover and mutation rate through evolution in DGA. For object detection, a CDM is constructed which is combined with the spatial segmentation result to produce the VOP.

To reduce the effect of noise and illumination variation for exact object detection, Su and Amer (2006) proposed a local/adaptive thresholding based moving object detection scheme. In this approach, each difference image is divided into a number of blocks and each block is tested for presence of region of change (ROC) with an ROC scatter estimation algorithm. The threshold value for each marked block containing ROC is averaged to obtain a global threshold value to segment the entire image frame. Here the difference image is obtained by taking pixel by pixel absolute difference in gray level of the reference frame (a frame in which no moving object is present) and the target frame (frame in which moving object is to be detected). Hence this approach is completely dependent on the presence of a reference frame.

It is evident from the previous literature that detection of moving objects from a video scene is always a challenging task. In this regard previous attempts are made by developing an *edge based compound Markov Random Field model* (Subudhi and Nanda, 2008a) for *spatial segmentation*, combination with a global thresholding scheme that provides a good output by preserving the object boundary with less computational time (Subudhi and Nanda, 2008b). The compound Markov Random Field model is modified (Subudhi and Nanda, 2008c) to detect slow moving video objects, and to reduce the effect of object-background misclassification. For further reduction of object-background misclassification, preliminary experiments of the proposed work are reported in (Subudhi et al., 2010).

In this article, we have designed a novel moving object detection scheme by combining two kinds of segmentation schemes: spatial and temporal. In global thresholding for temporal segmentation the threshold value for segmentation is obtained by thresholding the histogram of the entire difference image. This sometimes lead to misclassification of object and background pixels. Some pixels which actually belong to changed region (object region) may be identified as background, and vice versa. In this regard we propose a local thresholding scheme to segment the difference image by dividing it into a number of small regions/windows and each window is thresholded by adopting histogram thresholding method. In the proposed scheme the window/block size of the difference image is determined by measuring the entropy content of the considered window. The segmented regions from each window are combined to find the entire segmented image. This thresholded difference image is termed as the CDM and represents the changed regions corresponding to the moving objects in the given image frame. Here the difference image is generated by incorporating the label information of the pixels obtained by the spatially segmented image of two frames. In spatial segmentation we have used an MRF model for image modeling and the maximum a posteriori probability (MAP) estimate of the pixels are obtained by a combination of simulated annealing (SA) and iterated conditional mode (ICM) algorithms. It is observed that the entropy based adaptive window growing scheme gives better results towards moving object detection with less effect of object-background misclassification.

The results obtained by the proposed segmentation method are compared with those of JSEG (Deng and Manjunath, 2001), mean-shift (Comaniciu and Meer, 2002), and *MRF-edgeless* (Subudhi and Nanda, 2008a) methods of segmentation and is found to be better. Similarly, the VOPs generated by the proposed entropy based adaptive window selection scheme is compared with the VOPs obtained by the global thresholding approach (Subudhi et al., 2009), and is found to be better.

The organization of this article is as follows. In Section 2 the proposed moving object detection scheme is narrated with the help of a block diagram. In Section 3, spatial segmentation method using MRF framework is presented. In Section 4, temporal segmentation based on entropy based adaptive window scheme is given. Section 5 provides simulation results and analysis. Conclusion is presented in Section 6.

2. Proposed algorithm for moving object detection

A block diagrammatic representation of the proposed scheme is given in Fig. 1. We have used a combination of two types of segmentation schemes for moving object detection: spatial and temporal. Spatial segmentation helps in determining the boundaries of both still and moving objects in the scene, and temporal segmentation helps in determining the foreground and background parts of the scene.

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. The spatial segmentation in spatio-temporal framework has been cast as a pixel labeling problem. The pixel labels are estimated using MAP criterion. The MAP estimates of the pixel labels are obtained by a combination of both SA and ICM as in (Subudhi and Nanda, 2008a, 2009).

For temporal segmentation, we have proposed a local/adaptive thresholding scheme to threshold the difference image. In local thresholding we divide the difference image into a number of regions/windows and a histogram thresholding based scheme is used to segment each window. The difference image is obtained by incorporating the label information of the pixels obtained by the spatially segmented image of two frames termed as difference image. In the proposed local thresholding scheme the region/window size in the difference image is determined by the entropy content of the considered window. After the window size is determined, the histogram of the region is thresholded by Otsu's (1979) method. Then we combine the segmented regions from all windows to find the entire segmented image, called the CDM. The CDM represents the changed regions corresponding to the moving objects in the given image frame. The CDM is further modified based on the spatial and temporal segmentations of two image frames to construct the region corresponding to the moving object in the target frame. The thresholded image is fused with the spatial segmentation result of that frame to obtain the final temporal segmentation. Subsequently, the pixels corresponding to the foreground part of the temporal segmentation is used to display the VOP of that frame.

A schematic representation of the whole process is shown in Fig. 1. Frame *t* represents the observed image frame at *t*th instant of time. We model the *t*th frame with its 2nd 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 *t*th 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 *t*th frame. The whole process is performed in spatio-temporal framework. Temporal segmentation is obtained as follows. We obtain a difference image of two frames i.e., the *t*th and the $(t-d)$th frame and is thresholded by the proposed entropy based adaptive window selection scheme. The position of the object in the thresholded image represents the amount of movement performed by objects in the scene from the $(t-d)$th instant to the *t*th instant of time. The spatial segmentation result of the *t*th frame, the $(t-d)$th frame, along with VOP of the $(t-d)$th frame are used to perform a temporal segmentation of the *t*th frame. The pixels corresponding to the object regions of the temporal segmentation are replaced by the original pixels of the *t*th frame to obtain the VOP of the *t*th frame.

3. Spatial segmentation scheme

It is assumed that the observed video sequence y is a 3-D volume consisting of spatio-temporal image frames. y_t represents the video image frame at time t. Each pixel in y_t is a spatial site s

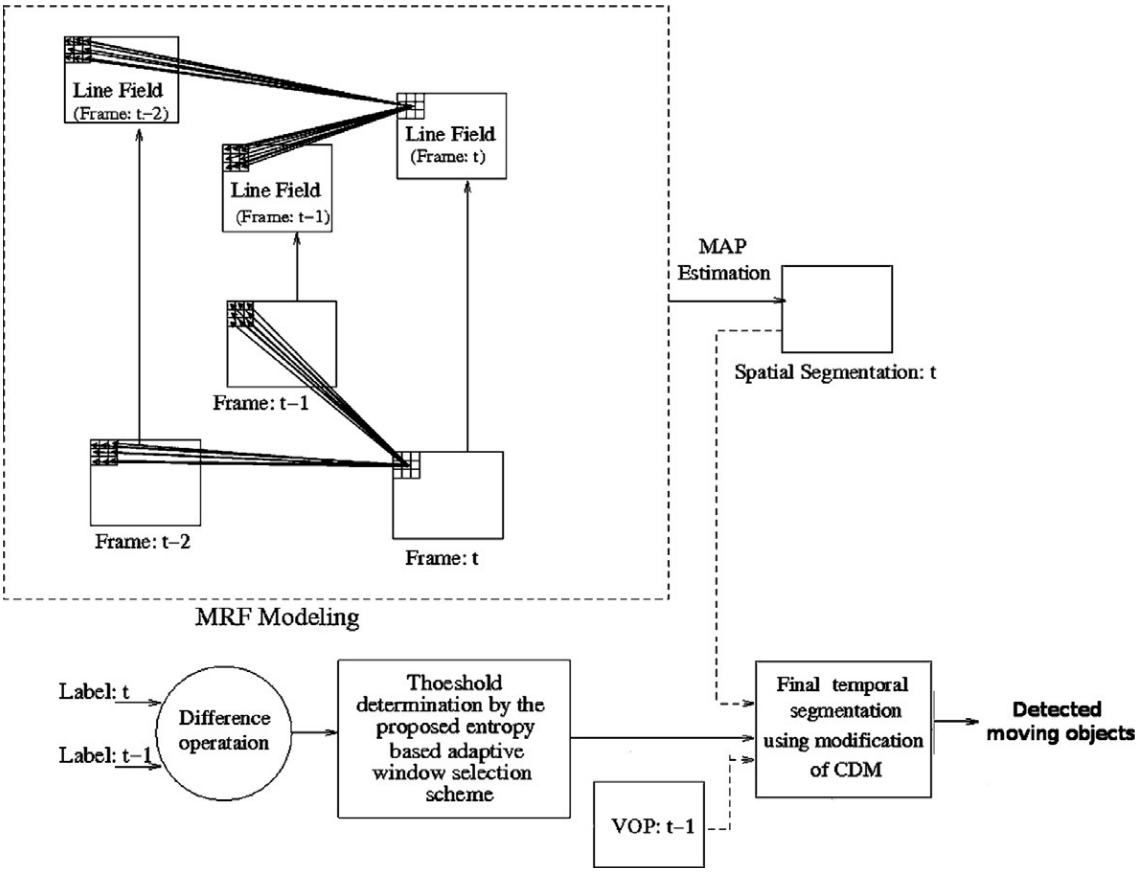


Fig. 1. Block diagram of the proposed moving object detection scheme.

denoted by y_{st}, q is a temporal site and e is an edge site. The set of all sites is represented by S. Then $S = \{s\} \cup \{q\} \cup \{e\}$. 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 the segmented version of y_t. Let us assume that X_t represents the MRF from which x_t is a realization. Similarly, the 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/edge map of the current frame x_t and the linefields/edgemaps of x_{t-1} and x_{t-2}.

The novelty of MRF model lies in the fact that it takes into account the spatial and temporal intrinsic characteristics of a region and is well known in the literature (Li, 2001). The regions are assumed to have uniform intensities. The considered image frame is assumed to come from an imperfect imaging modality. It is assumed that noise has corrupted the actual image to produce the observed image y_t. Given this observed image y_t we seek the image x_t, which maximizes the posterior probability. We denote the estimate of x_t by $\hat{x}_t$. Hence, it is assumed here that due to the presence of noise we are not able to observe x_t. But we can observe a noisy version of x_t as y_t. Noise is assumed to be white (additive i.i.d., i.e., independent and identically distributed). Hence Y_t can be expressed as (Li, 2001)

$$y_t = x_t + N(0, \sigma^2), \tag{1}$$

where $N(0,\sigma^2)$ is the i.i.d. noise with zero mean and σ^2 variance.

Fig. 2 shows a diagrammatic representation of the considered MRF modeling. Second order neighborhood is used here. For example, as shown in Fig. 2(a), (i,j)th pixel with its spatial neighbors are used to construct the second order clique function in spatial direction. Fig. 2(b) shows the considered 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, an 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 shown in Fig. 2(c).

The prior probability of the MRF framework can be represented by 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). The parameter T is the temperature constant and is considered to be $T = 1$ as in (Li, 2001). We have considered the following clique potential functions for the present work.

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

Analogously in the temporal direction,

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

and for the edgemap in the temporal direction

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

Here V_{sc} denotes spatial clique potential, V_{tec} denotes temporal clique potential and V_{teec} denotes temporal direction edge clique potential. α, β and γ are the parameters associated with the clique potential function. These are +*ve* constants and are determined by trial and error. We have used the additional features in the temporal direction and the whole model is referred to as *edgebased* model. Hence, in our a priori image model the clique potential function is a combination of the above three terms. Thus the energy function takes the following form

$$U(x_t) = \sum_{s,t} \{V_{sc}(x_{st}, x_{pt}) + V_{tec}(x_{st}, x_{qt}) + V_{teec}(x_{st}, x_{et})\}. \tag{2}$$

The observed image sequence y is assumed to be a degraded version of the actual image sequence x. The degradation process is assumed to be Gaussian. Thus, the label field x_t can be estimated from the observed random field Y_t by maximizing the following posterior probability distribution:

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

where $\hat{x}_t$ denotes the estimated label. The prior probability $P(Y_t = y_t)$ is constant and can be discarded for this purpose.

The prior probability $P(X_t = x_t)$ can be described as

$$P(X_t = x_t) = \frac{1}{z} e^{\frac{-U(x_t)}{T}} = \frac{1}{z} e^{-\frac{1}{T}\sum_{s,t}\{V_{sc}(x_{st},x_{pt}) + V_{tec}(x_{st},x_{qt}) + V_{teec}(x_{st},x_{et})\}}. \tag{4}$$

Assuming decorrelation among the three RGB planes for the color image and the variance to be the same among each plane (Perez, 1998; Kaiser, 2007), the likelihood function $P(Y_t = y_t | X_t = x_t)$ can be expressed as

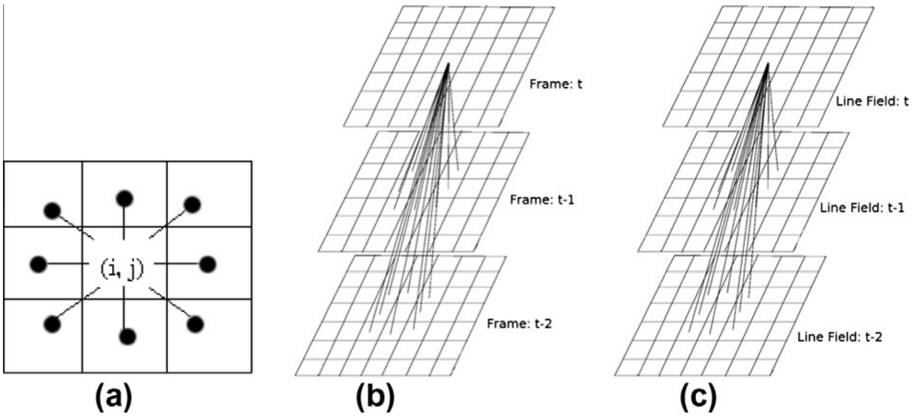


Fig. 2. (a) Neighborhood of a site for MRF modeling in the spatial direction, (b) MRF modeling taking two previous frames in the temporal direction, (c) MRF with two additional frames with line fields to take care of edge features.

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

Now putting Eqs. (4) and (5) in Eq. (3) we obtain

$$\hat{x_t} = \arg\min_{x_t} \left\{ \left[\frac{\|y_t - x_t\|^2}{2\sigma^2} \right] + \left[\sum_{s,t} \{ V_{sc}(x_{st}, x_{pt}) + V_{tec}(x_{st}, x_{qt}) + V_{teec}(x_{st}, x_{et}) \} \right] \right\}. \tag{6}$$

$\hat{x_t}$ is the MAP estimate and is obtained by a combination of SA and ICM algorithms as described in (Subudhi et al., 2009).

4. Temporal segmentation scheme

Let us consider two image frames at tth and $(t-d)$th instants of time. For temporal segmentation, we have obtained a difference image by considering the label information of the pixels from the spatial segmentation of two image frames. For difference image generation, at a particular pixel location, if the spatial segmentation of the two image frames are found to be the same (either object or background) then the difference image value is set to 0. Otherwise the pixel value in the difference image is obtained by taking a difference of the corresponding values of the R, G and B components of the considered frames. The use of such a difference image has an advantage of preserving the boundary information of the object. Usually the boundary of the object has a higher chance of misclassification. By considering a difference image with label information, these effect can be reduced. It is also found that the isolated pixels in the background regions may have large variation in gray value as compared to the other frames due to noise and illumination variation. Use of label difference image can also reduce this effects. Each channel of the obtained difference image is thresholded separately by the proposed entropy based window selection method. After obtaining the thresholded images for all the channels, they are fused by a logical *OR* operator. In a video, changes or movements of an object in the scene may reflect a change in a single color channel. An *OR* operator is very helpful in identifying those. A schematic representation of the above process is shown in Fig. 3, which shows the details of the “difference operation” and “threshold determination by the proposed entropy based adaptive window selection scheme” blocks of Fig. 1.

Determination of the threshold value for each channel of the label difference image is a very difficult task. If a global thresholding algorithm is used on the difference image: (i) a few pixels that actually correspond to the background in an image frame are identified as changed pixels, (ii) it also happens that a few pixels in the difference image that correspond to the actual changed regions and lie on the lower range of the histogram may be identified as unchanged pixels. An adaptive thresholding approach, where thresholding can be done on small parts of the image, can be used to overcome these problems. However, the choice of small parts or windows is a difficult issue. If a larger window is considered smaller changes cannot be detected properly; whereas in a small size window noise or pixels affected by illumination variation are more and may be detected as object pixels. In this regard we propose an entropy based adaptive window selection scheme to determine the block/window size. The threshold value for a particular window is obtained by Otsu’s scheme (Otsu, 1979). An union of all the thresholded blocks represents the CDM. The details of this scheme is given below.

4.1. Entropy based window selection for thresholding

The basic notion of window selection approach is to fix the window size primarily focussing on the information content of the window or the sub image and the whole image. In other words, fixing the size of the window depends on the entropy of the chosen window. In this approach initially an arbitrarily small window (in the present case 5 × 5) is considered (at the beginning of the image) and the entropy of the window is computed from the gray level distribution of this window and is denoted by H_w as

$$H_w = \sum_{i=1}^{L} p_i \log_e \left(\frac{1}{p_i} \right), \tag{7}$$

where p_i is the probability of occurrence of the ith gray level in the window and L is the maximum gray level. Local entropy is related to the variance of gray value of the window or the sub image. Entropy is more for a heterogeneous region, and less for a homogeneous region. Hence object-background transition regions will have more local entropy than that in non-transition regions of the image. It is observed that the information content of a heterogenous region/window in an image is close to some fraction of the entropy of the window or the sub image. If the entropy of the window is comparable to some fraction of the entropy of the whole image ($H_w > Th$, $Th = c * H_D$, Th is threshold, c is a constant in [0,1], and H_D is the entropy of the difference image D), that window is chosen for segmentation. Otherwise the window is incremented by Δw

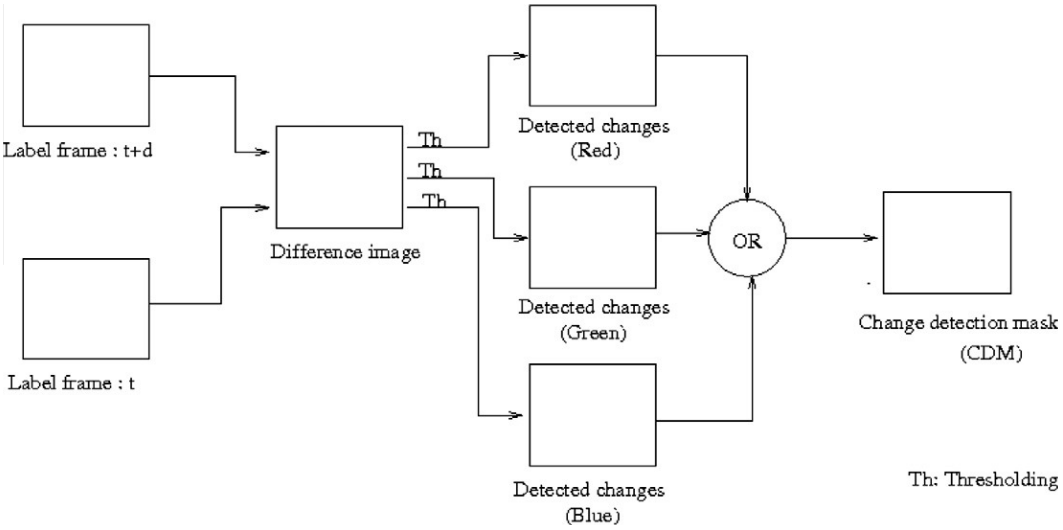


Fig. 3. Blockdiagram of temporal segmentation scheme.

(here it is considered 2) and the condition is tested again. Once a window has been selected for segmentation, the next window is selected from the rest of the difference image.

Fig. 4(a) illustrates the window growing method. First a window of size $n \times n$ is chosen and it is incremented with size Δw to make a window of size $(n + \Delta w) \times (n + \Delta w)$ and so on until the predefined condition $(H_w > Th)$ is satisfied. Fig. 4(b) shows that after fixing of a window, another window of size $n \times n$ is started from the adjacent side. In the final step, if the area of the remaining portion of the image is less than $n \times n$, then that area is taken as another window. Hence, there is no chance of overlapping between windows.

The final thresholded image is obtained by taking a union of all the considered thresholded window.

The salient steps of the proposed algorithm are enumerated below:

```
(i) Choose a window of size $w = n \times n$, where $n$ is a small positive integer.
(ii) Determine the entropy $H_w$ from the gray level distribution of the window $w$.
(iii) **if** $H_w > Th$ then
        • Fix the window size as $w$ and apply thresholding on it.
        • Set the region under the window as *covered*.
        • Start a new window of size $w = n \times n$ (adjacent to the previous window) from the *not yet covered* area of the image.
    **else**
        • Increase the window size by $\Delta w$, i.e., $w \leftarrow w + \Delta w$.
        • Repeat Steps 2–3 till the whole image is *covered*.
```

4.2. Temporal segmentation by modification of CDM

The CDM thus obtained by the proposed entropy based window selection scheme contains, either changed or unchanged (i.e., denoted as 1 or 0) pixels. This CDM also reflects the previous position of moving object (except the overlapped area) as a changed area, which is required to be eliminated in the current frame. To detect the moving object in the current frame, it requires a modification of the CDM so as to eliminate the object position in the previous frames and also to bring out the overlapped area corresponding to the moving object in the subsequent frame. To improve the temporal segmentation, obtained change detected output is combined with the VOP of the previous frame based on the label information of the current and the previous frames. The VOP of the previous frame ($(t-d)$th) is represented as a matrix of size $M \times N$ as

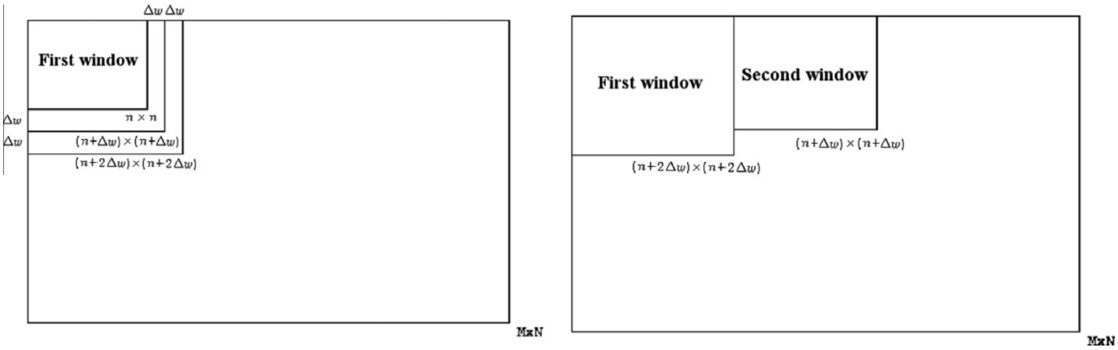


Fig. 4. Illustration for (a) window growing method, (b) starting of another window.

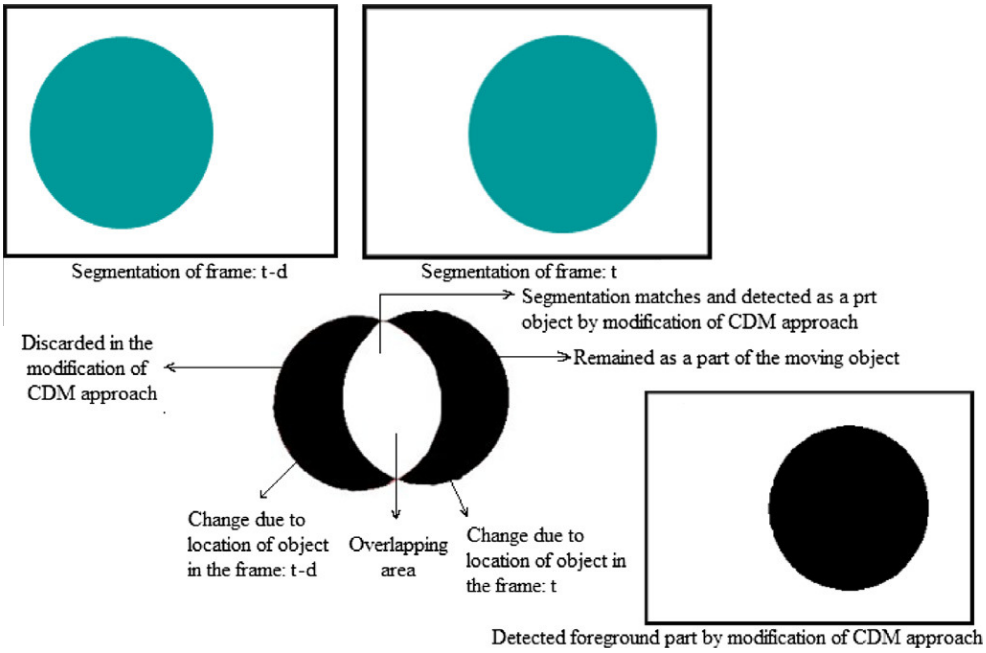
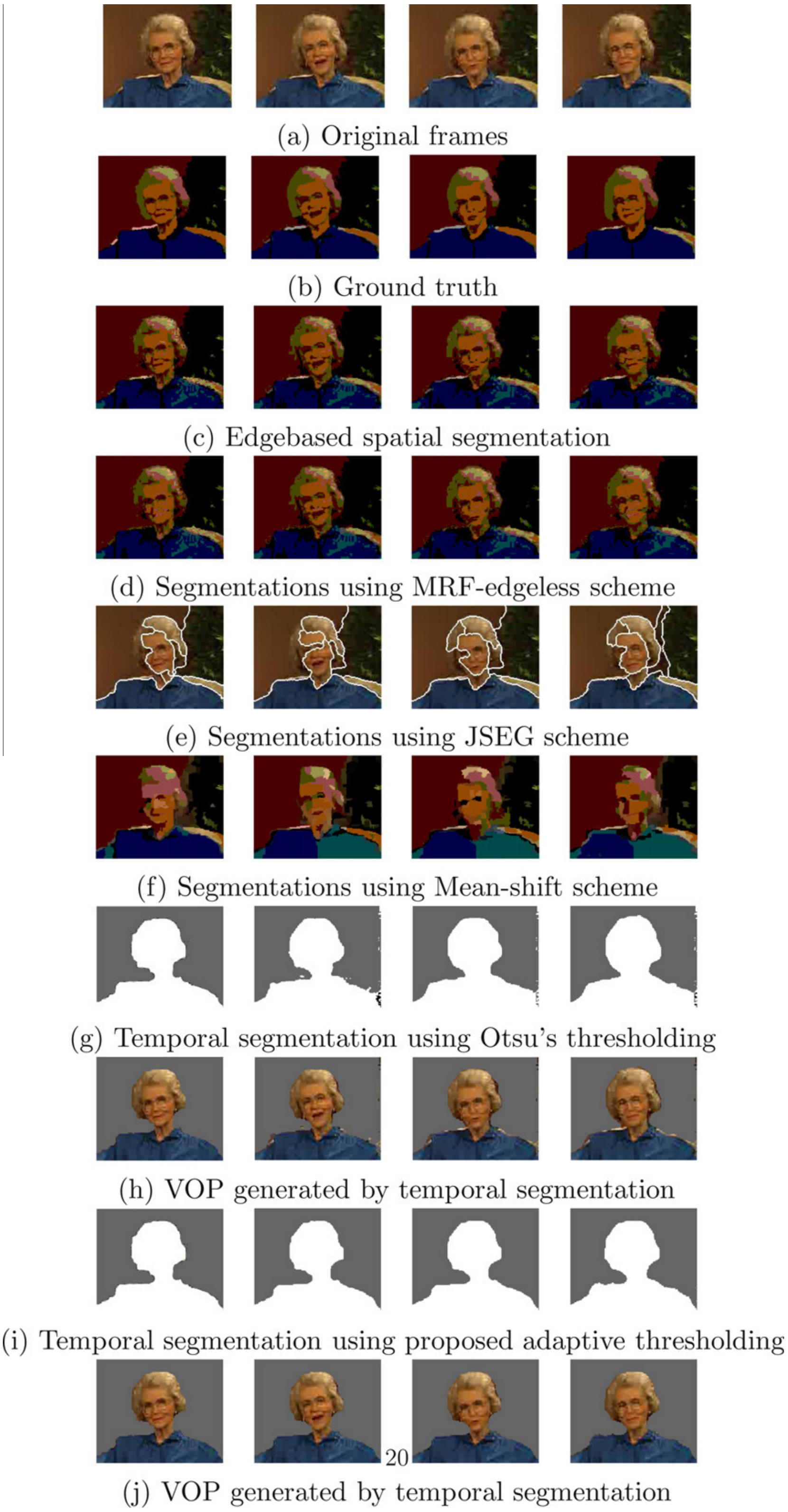

Segmentation of frame: t-d
Segmentation of frame: t
Segmentation matches and detected as a prt object by modification of CDM approach
Remained as a part of the moving object
Discarded in the modification of CDM approach
Change due to location of object in the frame: t-d
Overlapping area
Change due to location of object in the frame: t
Detected foreground part by modification of CDM approach

Fig. 5. Temporal segmentation by modification of CDM.

$$R(t-d) = \{r_{i,j}(t-d) | 0 \leqslant i \leqslant (M-1), 0 \leqslant j \leqslant (N-1)\},$$

where each element of the matrix i.e., $r_{i,j}(t-d)$ represents the value of the VOP at location (i,j) in $(t-d)$th frame. Here $R(t-d)$ is a matrix having the same size as that of the image frame (i.e., $M \times N$) and (i,j)th location represents the ith row and the jth column and is described as

$$r_{i,j}(t-d) = \begin{cases} 1, & \text{if it is in the object}, \\ 0, & \text{if it is in the background}. \end{cases}$$



(a) Original frames

(b) Ground truth

(c) Edgebased spatial segmentation

(d) Segmentations using MRF-edgeless scheme

(e) Segmentations using JSEG scheme

(f) Segmentations using Mean-shift scheme

(g) Temporal segmentation using Otsu's thresholding

(h) VOP generated by temporal segmentation

(i) Temporal segmentation using proposed adaptive thresholding

(j) VOP generated by temporal segmentation

Fig. 6. VOP generated for Grandma video sequence for frames (12th, 37th, 62nd, 87th).

If a pixel is found to have $r_{i,j}(t-d)=1$, it is a part of the moving object of the previous frame $(t-d)$; otherwise it belongs to the background of the previous frame. Based on this information, CDM is updated as follows. If a pixel at location (i,j) in the current frame (at time t) belongs to a moving object in the previous frame (i.e., $r_{i,j}(t-d)=1$) and its label obtained by spatial segmentation scheme in tth frame is the same as the corresponding pixel in the previous frame ($(t-d)$th), the pixel is marked as the foreground area in the current frame else as a background. The modified CDM thus represents the temporal segmentation.

Let us consider that a disc is moving in the horizontal direction. Fig. 5 represents the complete process of temporal segmentation by modification of CDM. The CDM obtained by label difference represents two regions: one on the left side due to the location of the object in the $(t-d)$th frame; and the second region on the right side represents the changes detected in the tth frame. In the overlapping region of these two no change will be detected. By considering the spatial segmentation of both the frames, the region on the left side of the CDM is eliminated as the segmentation of the pixels in this region of the two frames ($(t-d)$ and t) does not match with each other. Similarly, in the overlapping area, the segmented output of all pixels matches with each other. Hence, it is detected as a part of the moving object. The final (temporal segmentation) output represents the complete moving object.

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.

5. Simulation results and discussion

To establish the effectiveness of the proposed scheme, we have tested it on three different video sequences: *Grandma*, *Canada traffic*, *Karlsruhe taxi-2*. Since changes in between the consecutive frames are very less, we have considered frames after a particular interval of time where a reasonable amount of change is expected to have occurred. To provide a quantitative evaluation of the proposed scheme, we have provided two *ground-truth* based performance measures. One measure is considered for quantitative evaluation of the proposed spatial segmentation scheme and another measure is for quantitative evaluation of the obtained moving object locations. For evaluating the accuracy of the spatial segmentation, we have used the pixel by pixel comparison of the *ground-truth* image with the obtained spatial segmentation results. This measure is also called *number of misclassified pixels*. To evaluate the performance of the moving object detection, we have considered the precision and recall measures. It may be noted that for a better spatial segmentation the *number of misclassified pixels* should be less. Similarly, for a better object detection, the precision and recall measure should be more.

The 1st example considered, consists of 12th, 37th, 62nd and 87th frames of the *Grandma* sequence having a single moving object. Here the moving object is Grandma. Fig. 6(a) and (b) show the original and the manually constructed ground-truth images (spatial segmentation) of 12th, 37th, 62nd and 87th frames of *Grandma* sequence. The edge based compound MRF model is used for modeling the attributes of these image frames and the corresponding MAP estimate is obtained by a combination of SA and ICM algorithms. The edgebased spatial segmentation results

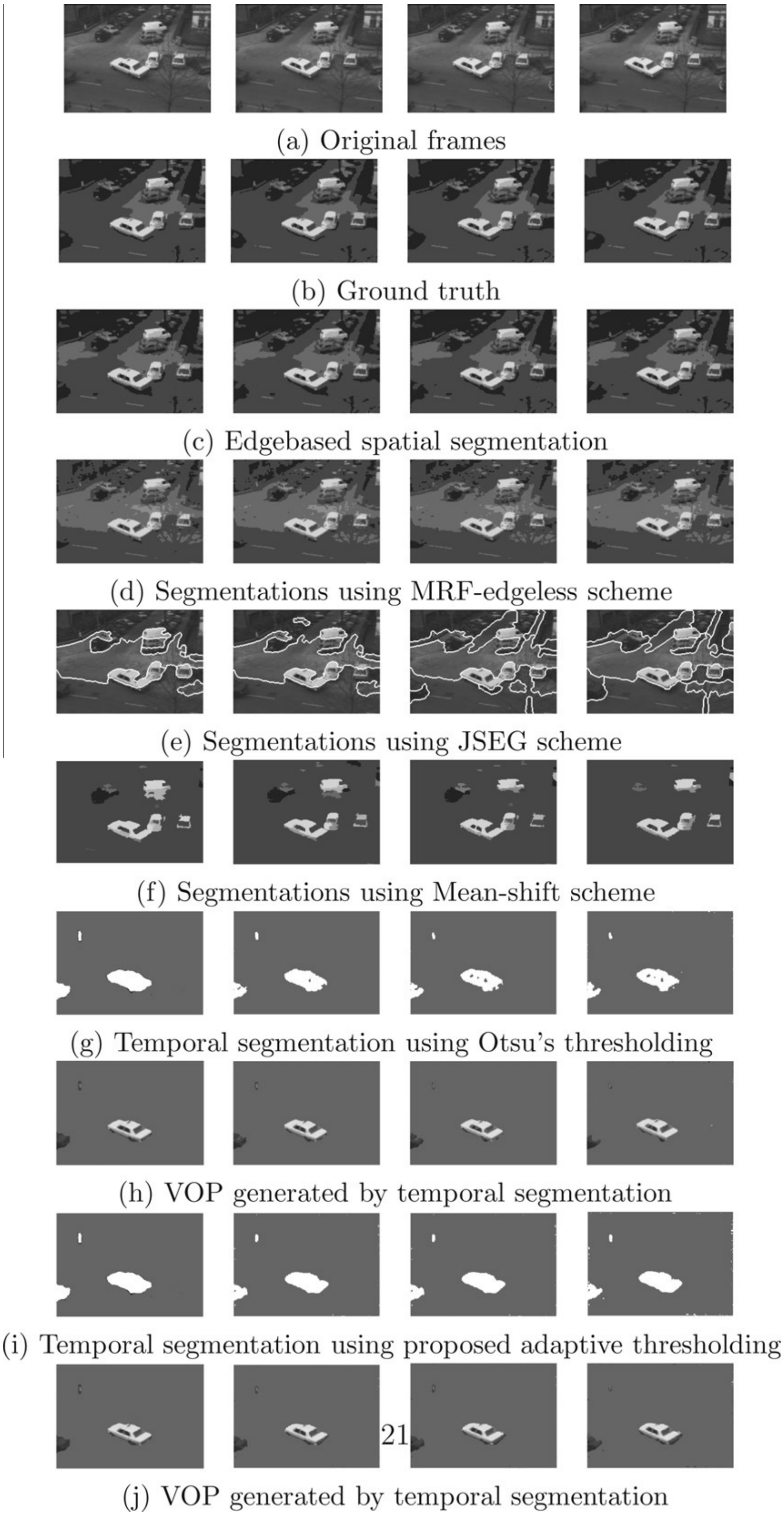
Table 1
Number of misclassified pixels.

Video	Frame no.	Edgeless	Proposed	JSEG	Mean shift
Grandma	12	231	**181**	5185	812
	37	380	**114**	3565	1069
	62	379	**107**	3422	934
	87	304	**90**	2950	878
	3	720	**115**	1936	2735
Canada traffic	4	951	**470**	3123	3440
	5	1845	**593**	2523	3523
	6	937	**524**	1500	2700
	3	841	**360**	1179	1155
Karlsruhe taxi-2	4	731	**317**	903	1100
	5	766	**337**	1155	1031
	6	680	**215**	1302	917

Table 2
Precision and recall count.

	Frame number	Proposed		Otsu's thresholding	
		Precision	Recall	Precision	Recall
Grandma	12	0.97	0.95	0.97	0.95
	37	0.98	0.93	0.90	0.89
	62	0.98	0.97	0.90	0.89
	87	0.96	0.95	0.90	0.89
Canada traffic	3	0.90	0.98	0.90	0.98
	4	0.93	0.91	0.84	0.85
	5	0.88	0.81	0.85	0.76
	6	0.90	0.93	0.82	0.74
Karlsruhe taxi-2	3	0.90	0.96	0.90	0.96
	4	0.97	0.93	0.88	0.81
	5	0.90	0.89	0.83	0.82
	6	0.89	0.84	0.79	0.77

obtained for these frames are shown in Fig. 6(c). The MRF model parameters used for *Grandma* sequence are $\alpha = 0.05$, $\beta = 0.009$, $\gamma = 0.007$ and $\sigma = 5.19$. The spatial segmentation result obtained for these frames are compared with those obtained with MRF-edgeless (Subudhi et al., 2009), JSEG (Deng and Manjunath, 2001) and mean-shift (Comaniciu and Meer, 2002) based segmentation schemes. The spatial segmentation results obtained for these frames with MRF-edgeless approach of segmentation scheme is displayed in Fig. 6(d), where it is found that nose, mouth, and spectacle of Grandma are merged with her face region. Considering the


(a) Original frames

(b) Ground truth

(c) Edgebased spatial segmentation

(d) Segmentations using MRF-edgeless scheme

(e) Segmentations using JSEG scheme

(f) Segmentations using Mean-shift scheme

(g) Temporal segmentation using Otsu's thresholding

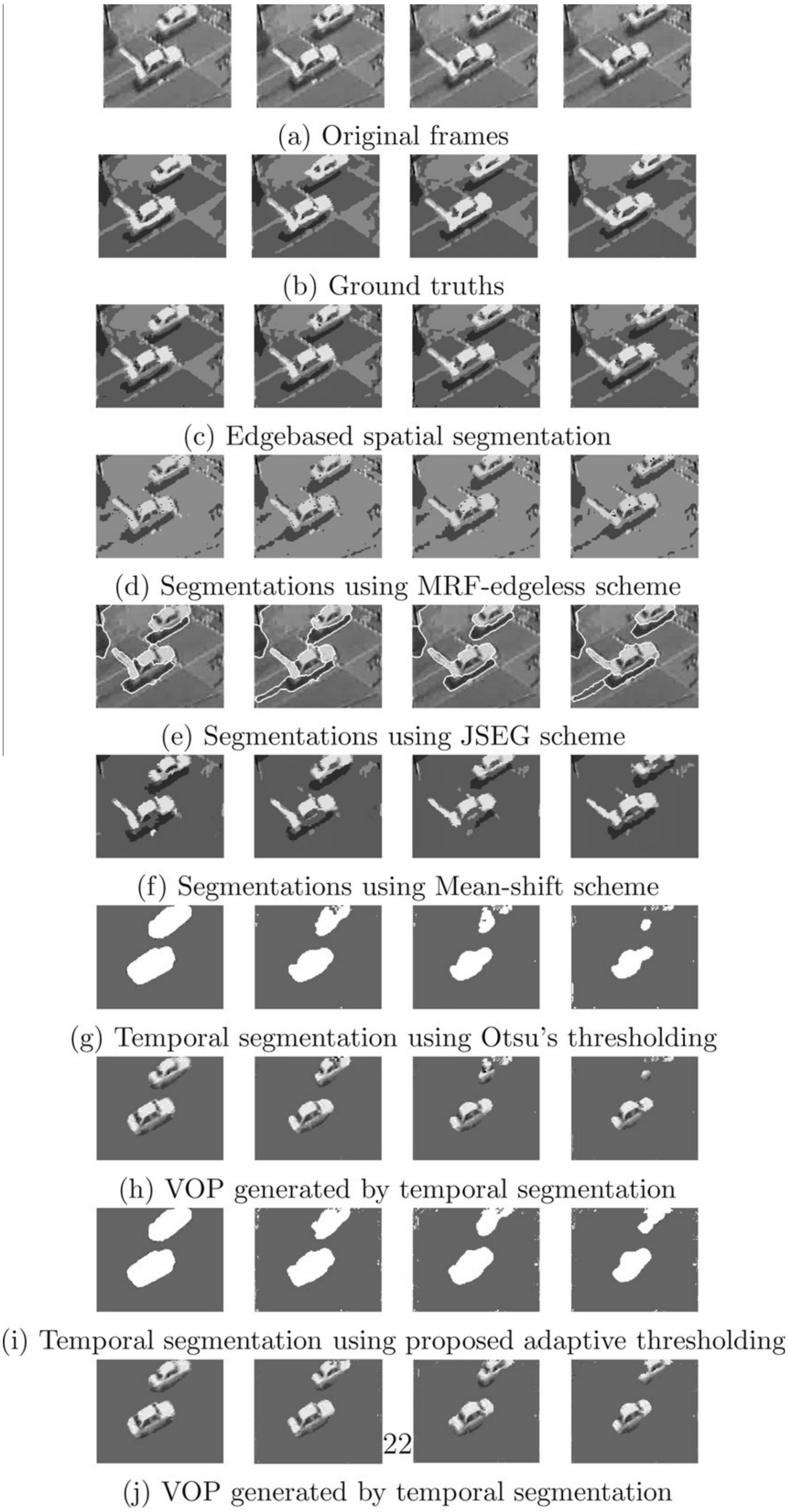
(h) VOP generated by temporal segmentation

(i) Temporal segmentation using proposed adaptive thresholding

(j) VOP generated by temporal segmentation

Fig. 7. VOP Generated for *Canada traffic* video sequence for frames (3rd, 4th, 5th, 6th).

results of JSEG (shown in Fig. 6(e)) it is found that the regions like hair, collar, lip, eyes, nose, etc. of the Grandma are merged into a single class. Few regions like hair, left eye of Grandma are merged into background. Similarly, the results obtained by mean-shift scheme are shown in Fig. 6(f). It is found from these results that a better segmentation is obtained than JSEG scheme but few regions in the grandma are merged or missed. However, the minute edge details in the image frames are reflected in case of edgebased spatial segmentation approach. By comparing these results with a set of manually constructed ground-truth images, it is observed


(a) Original frames

(b) Ground truths

(c) Edgebased spatial segmentation

(d) Segmentations using MRF-edgeless scheme

(e) Segmentations using JSEG scheme

(f) Segmentations using Mean-shift scheme

(g) Temporal segmentation using Otsu's thresholding

(h) VOP generated by temporal segmentation

(i) Temporal segmentation using proposed adaptive thresholding

(j) VOP generated by temporal segmentation

Fig. 8. VOP Generated of *Karlsruhe taxi-2* video sequence for frames (3rd, 4th, 5th, 6th).

that the number of misclassified pixels is quite less in case of edge-based spatial segmentation approach than that of MRF-edgeless and JSEG approach of segmentation as shown in Table 1. The temporal segmentation is obtained using an original frame difference followed by Otsu's (1979) approach and the result is shown in Fig. 6(g). The corresponding detected objects are shown in Fig. 6(h). It is observed that some background pixels are treated as foreground and some noisy pixels are still present with the foreground. Hence, there exists an effect of object and background misclassification. The temporal segmentation results obtained by the proposed adaptive window based scheme is shown in Fig. 6(i), where it can be observed that the object background misclassification is quite less and the corresponding detected objects are shown in Fig. 6(j). The precision and recall count of two object detection schemes are provided in Table 2.

The next video considered is *Canada traffic* sequence with three objects and each one is having a different speed. This video contains moving objects like: a black car, a white car and a person (moving in lawn). Fig. 7(a) shows the 3rd, 4th, 5th and 6th original image frames of this sequence. Corresponding manually constructed spatial segmentation ground-truth images are shown in Fig. 7(b). The spatial segmentation results of these frames are obtained by edgebased compound MRF model and are shown in Fig. 7(c). The MRF model Parameters used for *Canada Traffic* video sequence are $\alpha = 0.01$, $\beta = 0.009$, $\gamma = 0.007$ and $\sigma = 3.0$. The spatial segmentation result of these frames using edgeless-MRF scheme are displayed in Fig. 7(d). It is observed from these results that few portions of the moving black car and the person in the lawn are merged into background. It is also observed that a few of the still objects are also merged with the background. The JSEG based spatial segmentation of these frames are displayed in Fig. 7(e). These results show that the person and the black car also got merged into the background. Similarly, the result obtained by mean-shift scheme are shown in Fig. 7(f). These results show that the black car and the person moving in the lawn are merged into background. Some parts of the white car is also merged into the background. The misclassification error obtained with different spatial segmentation schemes for these frames are provided in Table 1. It is found from Fig. 7(g) and (h) that the global thresholding approach is not able to detect properly two objects (black car and the man). Similarly, the rear end of the white car is also not detected properly. Hence there are object and background misclassifications. As observed from Fig. 7(i) and (j), the two objects (black car and the man) along with the white car are detected properly using adaptive thresholding approach. Corresponding precision and recall values are put in Table 2.

The last video sequence we have considered is *Karlsruhe taxi-2* sequence. Fig. 8(a) shows the 3rd, 4th, 5th and 6th frames of *Karlsruhe taxi-2* sequence with two moving objects. This is also an illumination variate noisy sequence. The considered MRF model parameters for *Karlsruhe taxi-2* sequence are $\alpha = 0.01$, $\beta = 0.008$, $\gamma = 0.007$ and $\sigma = 4.0$. The edgebased spatial segmentation results of these frames are shown in Fig. 8(c). Corresponding results using MRF-edgeless, JSEG and mean-shift schemes are shown in Fig. 8(d)–(f). Otsu's global thresholding approach produces results with many missing part of moving objects and are shown in Fig. 8(g) and (h). Fig. 8(i) shows the temporal segmentation results obtained using the proposed adaptive window selection scheme. It can be seen from these sequences that all parts of the moving objects have been detected with less (object background) misclassification (shown in Fig. 8(j)). It is found that using Otsu's global thresholding scheme, one of the moving cars is almost missed whereas the proposed adaptive window based scheme provided better results. The precision and recall value for this sequence are put in Table 2.

From the above experiments, we observed that the precision and recall value is more for the proposed scheme than the non window based global thresholding scheme. Hence, we may conclude that the VOPs obtained by the proposed entropy based adaptive window based thresholding scheme provide better results for moving object detection than those obtained by non window based global thresholding scheme. Hence a less effect of object background misclassification error is noticed.

The proposed scheme is implemented in a *Pentium*4(*D*), 3 GHz, *L*2 *cache* 4 MB, 1 GB *RAM*, 667 *FSB* PC with *Fedora* – *Core* operating system and *C* programming language.

6. Conclusion

In this article we address the problem of moving object detection. The proposed technique uses a combination of two segmentation schemes: temporal and spatial. For temporal segmentation, we have proposed a local/adaptive thresholding scheme to segment the difference image into object and background. The difference image is divided into a number of regions/windows, and each region is thresholded by a histogram thresholding approach. In the proposed scheme window size is determined by measuring the entropy content of the considered window. The regions corresponding to each thresholded window of the difference image are combined to form the change detection mask (CDM). For temporal segmentation we have used a label difference image opposed to an original image frame difference. Here the label difference image is generated by taking the label information of the pixel from the spatially segmented output of two image frames. The spatial segmentation of an image frame is obtained by modeling both spatial and temporal attributes of image frames with a compound MRF model. Corresponding MAP estimate is obtained by a combination of simulated annealing and iterated conditional mode algorithms. It is observed that this approach gives better result towards moving object detection with less effects of object background misclassification as compared to the non-window based global thresholding scheme.

In the present work MRF model parameters are set manually. Our future work will focus on estimation of MRF model parameters using some sort of parameter estimation scheme. We are also looking at related problems with videos captured by moving camera where existing approach does not yield good results.

Acknowledgment

The authors like to acknowledge the through and constructive comments provided by the reviewers and the editors on this paper. Authors would also like to thank the Department of Science and Technology, Government of India and University of Trento, Italy, the sponsors of the ITPAR program.

References

Beleznai, C., Frnhstnck, B., Bischof, H., 2006. Human tracking by fast mean shift mode seeking. J. Multimedia 1 (1), 1–8.

Chuang, C.H., Hsieh, J.W., Tsai, L.W., Chen, S.Y., Fan, K.C., 2009. Carried object detection using ratio histogram and its application to suspicious event analysis. IEEE Trans. Circuits Systems Video Technol. 19 (6), 911–916.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell. 24 (5), 603–619.

Deng, Y., Manjunath, B.S., 2001. Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. Pattern Anal. Machine Intell. 23 (8), 800–810.

Forsyth, D.A., Ponce, J., 2003. Computer Vision a Modern Approach. Prentice Hall, New Jersey.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. 6 (6), 564–584.

Gonzalez, R.F., Woods, R.E., 2001. Digital Image Processing. Pearson Education, Singapore.

Haralick, R.M., Shapiro, L.G., 1992. Computer and Robot Vision. Addison-Wesley Publishing Company, New York.

Hinds, R.O., Pappas, T.N., 1995. An adaptive clustering algorithm for segmentation of video sequences. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, vol. 4, pp. 2427–2430.

Huang, P.S., Harris, C.J., Nixon, M.S., 1999. Human gait recognition in canonical space using temporal templates. IEE Proc. Vision Image Signal Process. 146 (2), 93–102.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems Man Cybernet. Part C 34 (3), 334–352.

Hwang, S.W., Kim, E.Y., Park, S.H., Kim, H.J., 2001. Object extraction and tracking using genetic algorithms. In: Proc. Internat. Conf. on Image Processing, vol. 2. Thessaloniki, Greece, pp. 383–386.

Kaiser, M.S., 2007. Statistical dependence in Markov random field models, Preprint 2007-1, Department of Statistics, Iowa State University, Ames, Iowa.

Kim, E.Y., Park, S.H., 2006. Automatic video segmentation using genetic algorithms. Pattern Recognition Lett. 27 (11), 1252–1265.

Kim, M., Choi, J., Kim, D., Lee, H., 1999. A VOP generation tool: Automatic segmentation of moving objects in image sequences based on spatio-temporal information. IEEE Trans. Circuits Systems Video Technol. 9 (8), 1216–1226.

Li, S.Z., 2001. Markov Random Field Modeling in Image Analysis. Springer, Japan.

Makris, D., Ellis, T., 2002. Path detection in video surveillance. Image Vision Comput. 20 (12), 895–903.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. System Man Cybernet. 9 (1), 62–66.

Perez, P., 1998. Markov random fields and images. CWI Q. 11 (4), 413–437.

Satake, J., Miura, J., 2009. Robust stereo-based person detection and tracking for a person following robot. In: Proc. IEEE Internat. Conf. on Robotics and Automation, pp. 1–6.

Schiele, B., Andriluka, M., Majer, N., Roth, S., Wojek, C., 2009. Visual people detection: Different models, comparison and discussion. In: Proc. IEEE Internat. Conf. on Robotics and Automation, pp. 1–8.

Shi, Q., Wang, L., Chen, L., Smola, A., 2008. Discriminative human segmentation and recognition using semi-Markov model. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. In: Proc. Internat. Conf. on Computer Vision and Pattern Recognition, pp. 2246–2252.

Su, C., Amer, A., 2006. A real time adaptive thresholding for video change detection. In: Proc. IEEE Internat. Conf. on Image Processing, pp. 157–160.

Subudhi, B.N., Nanda, P.K., 2008a. Compound Markov random field model based video segmentation. In: Proc. SPIT-IEEE R10 Colloquium Internat. Conf., vol. 1, pp. 97–102.

Subudhi, B.N., Nanda, P.K., 2008b. Moving object detection using compound Markov random field model. In: Proc. IEEE National Conf. Computational Intelligence, Control and Computer Vision in Robotics and Automation, vol. 1, pp. 198–204.

Subudhi, B.N., Nanda, P.K., 2008c. Detection of slow moving object using compound Markov random field model. In: Proc. IEEE TENCON, vol. 1, pp. 1–6.

Subudhi, B.N., Nanda, P.K., Ghosh, A., 2010. Moving object detection using MRF model and entropy based adaptive thresholding. In: Proc. IEEE 2nd Internat. Conf. on Human Computer Interaction. Springer, pp. 155–161.

Subudhi, B.N., Nanda, P.K., Ghosh, A., 2011. A Change information based fast algorithm for video object detection and tracking. IEEE Trans. Circuits Systems Video Technol. 21 (7), 993–1004.

Tekalp, A.M., 1995. Digital Video Processing. Prentice Hall, New Jersey.

Veeraraghavan, A., Roy-Chowdhury, A.K., Chellappa, R., 2005. Matching shape sequences in video with applications in human movement analysis. IEEE Trans. Pattern Anal. Machine Intell. 27 (12), 1896–1909.

Verma, V., Gordon, G., Simmons, R., Thrun, S., 2004. Particle filters for rover fault diagnosis. Rob. Autom. Mag., 54–64, special issue on Human Centered Robotics Dependability.

Yong, W., Bhandarkar, S.M., Kang, L., 2007. Semantics-based video indexing using a stochastic modeling approach. In: Proc. IEEE Internat. Conf. on Image Processing, vol. 4, pp. 313–316.

Zhang, Y.J., 2006. Advances in Image and Video Segmentation. IRM Press, New York.

Zucker, S.W., 1976. Region growing: Childhood and adolescence. Comput. Graphics Image Process. 5, 382–399.