Object Detection From Videos Captured by Moving Camera by Fuzzy Edge Incorporated Markov Random Field and Local Histogram Matching

Ashish Ghosh, *Member, IEEE,* Badri Narayan Subudhi, *Student Member, IEEE,* and Susmita Ghosh

***Abstract*—In this paper, we put forward a novel region matching-based motion estimation scheme to detect objects with accurate boundaries from videos captured by moving camera. Here, a fuzzy edge incorporated Markov random field (MRF) model is considered for spatial segmentation. The algorithm is able to identify even the blurred boundaries of objects in a scene. Expectation Maximization algorithm is used to estimate the MRF model parameters. To reduce the complexity of searching, a new scheme is proposed to get a rough idea of maximum possible shift of objects from one frame to another by finding the amount of shift in positions of the centroid. We propose a χ^2-test-based local histogram matching scheme for detecting moving objects from complex scenes from low illumination environment and objects that change size from one frame to another. The proposed scheme is successfully applied for detecting moving objects from video sequences captured in both real-life and controlled environments. It is also noticed that the proposed scheme provides better results with less object background misclassification as compared to existing techniques.**

***Index Terms*—Edge analysis, fuzzy sets, image segmentation, maximum *a posteriori* probability estimation, motion analysis.**

I. Introduction

EMULATION of human visual perception capabilities by machines is a challenging problem in computer vision. It deals with algorithms or theory of building artificial vision systems that extract information about the 3-D structure, orientation, spatial property of a generic object in a given scene from 2-D images [1]. This information analysis leads to provide a better way of detecting moving objects at different instants of time in a given video sequence. Moving object detection in a scene has wide applications, such as visual surveillance [2], activity recognition [3], robotics [4], and so on.

Moving object detection in a given video scene can be achieved by two different ways: 1) motion detection/change detection; and 2) motion estimation. Change or motion detection is the process of identifying the moving objects by finding the changed and unchanged regions from the extracted image frames when the camera is fixed and the objects are moving. For motion estimation, the motion vectors are computed 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 [1]. It is very difficult to detect a moving object from a video captured with moving camera, where objects and the camera both may have movements. To solve such problems one can adhere to the approaches, such as optical flow scheme [1].

An optical flow-based motion vector computation for moving object detection is well studied by Lucas and Kanade [5] to track the translational motion of an object in the scene, popularly known as KLT tracker [5]. However, this scheme is unable to give good results in objects with complex motions and dynamic background. A combination of spatial segmentation and motion prediction/estimation is proved to be a better approach toward this [6], [7].

An early work for object detection by a computationally efficient watershed-based segmentation scheme is suggested by Wang [6], where watershed algorithm is used for object boundary identification. This method has two main drawbacks: 1) the complexity is high; and 2) spatial segmentation produces an over-segmented result. To enhance the accuracy, Deng and Manjunath [7] proposed a robust moving object detection scheme, where the spatial segmentation of an image frame is obtained by color quantization followed by edge preserving region growing. For temporal segmentation, the regions corresponding to objects are matched in the temporal direction by computing the motion vectors of the object regions in the target frame. It is considered that the use of region-based spatial segmentation scheme cannot handle the spatial ambiguities of image gray values; and hence produces an over-segmented result that leads to misclassification in moving object detection.

The gray level of pixels in a video with high uncertainty and high ambiguity makes it difficult to detect moving objects with reasonable accuracy by nonstatistical spatial segmentation schemes. 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 [8], [9], in this context, is proved to be

Manuscript received August 2, 2011; revised October 31, 2011; accepted December 2, 2011. Date of publication March 9, 2012; date of current version July 31, 2012. A part of this work was done when A. Ghosh visited RWTH, Aachen University, Aachen, Germany, in September 2010. The work of B. N. Subudhi was supported by the Council of Scientific and Industrial Research, under Senior Research Fellowship 9/93 (0137)/2011. This paper was recommended by Associate Editor B. Zeng.

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

S. Ghosh is with the Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India (e-mail: susmitaghoshju@gmail.com).

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

a better framework. A robust work on MRF-based object detection was demonstrated by Kuo *et al.* [10], where a combination of temporal and spatial constraints of the image frames are used with MRF to obtain the moving object location from one frame to another. Some similar works were also reported in [11] to demonstrate the effectiveness of the MRF model in moving object detection from complex video scenes. However, the schemes [8] and [9] are limited to the video sequences captured by a fixed camera; and [11] is limited to passive millimeter wave images. As per discussion by Babacan and Pappas in their earlier work [12], it is observed that the use of multilayer MRF model (introduction of temporal cliques) for segmentation preserves temporal continuity, but can cause fragmentation (segment one object in multiple parts).

Jodoin *et al.* [13] proposed a robust moving object detection and tracking scheme for both fixed and moving camera captured video sequences, where MRF is used for label fields fusion. Recently, the capability of the MRF model was well exploited by Wang [14] for detecting moving vehicles in different weather conditions. However, this approach is limited due to its only applicability in gray scale videos.

In order to handle the spatial ambiguities of gray values, in this paper, we propose a scheme that is able to detect moving objects with accurate boundaries from videos captured by moving camera. This algorithm uses a region-based motion estimation scheme. Here, we have proposed a new kind of MRF-MAP framework, where fuzzy edge strength at each pixel location is incorporated in the MRF modeling. The scheme is able to preserve the object boundary for segmentation. RGB color features are used. The spatial segmentation problem is solved using the MAP estimation principle. The parameters of MRF are estimated by Expectation Maximization (EM) algorithm. In a region-based motion estimation scheme, to reduce the complexity of searching, a rough knowledge of maximum possible shift in object from one frame to another is obtained by calculating the amount of shift in the centroid of the object from one frame to another. Moving objects in the target frame is detected by χ^2-test-based local histogram matching, which helps to detect objects present in a low illumination environment and may change its size from one frame to another.

The effectiveness of the proposed fuzzy edge map incorporated MRF modeling for spatial segmentation is evaluated by comparing the results obtained by it with those of the spatial segmentation by meanshift [15], conventional MRF modeling [13], and MRF modeling with deterministic edge kernel. It is found that the proposed spatial segmentation scheme provides good results as compared to other approaches. Similarly, the moving object detection results obtained by the proposed scheme are compared with those obtained by optical flow [5], label fusion [13], and level set [16] based schemes, and are found to have less object-background misclassification. Accuracy of the proposed object detection scheme was evaluated by Erdem and Sankur's performance evaluation measures [17].

II. Proposed Algorithm for Object Detection

A block diagrammatic representation of the proposed scheme is shown in Fig. 1. The proposed scheme uses a

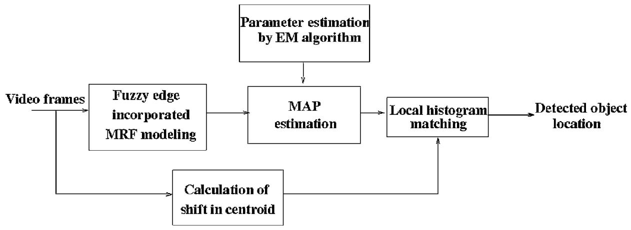


Fig. 1. Block diagram of the proposed scheme.

region based motion estimation scheme for detecting moving objects from a given video sequence. In the proposed scheme, the incoming image frames of the given video sequence are initially segmented into a number of homogenous regions by fuzzy edge map incorporated MRF modeling. In this scheme, the fuzzy edge strength of each pixel is incorporated in the MRF modeling that is able even to preserve the blurred boundary in the segmentation. In this context, we mention that fuzzy set theories are reputed to handle uncertainties to a reasonable extent, arising from deficiencies of information available from a situation (the deficiency may result from incomplete, ill-defined, not fully reliable, vague and contradictory information). Thus, we use the concept of fuzzy-edge modeling to handle spatial ambiguities at the object boundaries.

The spatial segmentation problem is solved using the MAP estimation principle. The parameters of MRF are estimated by EM algorithm.

In the subsequent stages of the proposed scheme, region-based motion estimation is used to find the moving objects in the considered video image frame. In the region-based motion estimation scheme, to reduce the complexity of searching, a rough knowledge of maximum possible shift in object from one frame to another is obtained by calculating the amount of shift in the centroid of the object from one frame to another. The moving objects in the target frame of a given video are detected by χ^2-test-based local histogram matching. The union of all matched regions represents the moving objects in the target frame.

III. Spatial Segmentation Using MRF Model

In this section, we discuss the process of MRF-based image modeling, MRF-MAP estimation, and EM algorithm for MRF model parameter estimation.

A. Spatial MRF 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. Here, we assume that the observed image frame $y_t = y_t(i, j)$ is a spatio-contextual entity of size $M \times N$. Each pixel in y_t is assumed as a site s denoted by y_{st}. Thus, y_{st} denotes a spatio-temporal coordinate 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. Here, $X_t = X_t(i, j)$ is discrete valued and can take values from $Q = \{q_1, q_2, ..., q_m\}$ for each $(i, j) \epsilon L$, where L is a finite lattice of size $M \times N$. A realization of $X_t = x_t$ is a partitioning of the lattice into m region types such that $x_t(i, j) = q_k$, if pixel (i, j) belongs to the kth region type. Each region type can occur in more than one location in the lattice.

B. Proposed Fuzzy Edge Incorporated MRF Modeling

In spatial domain, X_t represents the MRF model of x_t and using Hamersely Clifford's theorem [19] the prior probability can be expressed as Gibb's distribution with $P(X_t) = \frac{1}{z} e^{\frac{-\bar{U}(x_t)}{T}}$, where z is the partition function expressed as $z = \sum_{x_t} e^{\frac{-\bar{U}(x_t)}{T}}$, $\bar{U}(x_t)$ is the energy function, a function of clique potentials $V_c(x_t)$ [19]. This can be given as $\bar{U}(x_t) = \sum_{c \epsilon C} V_c(x_t)$.

According to Pott's model [20] the clique potential function $V_c(x_t) = -\alpha$ if all labels in possible set of cliques (C) are equal; otherwise, $V_c(x_t) = +\alpha$. Equal penalization of all the boundary pixels results in a greater penalty to weak edge pixels and a lesser penalty to the strong edge pixels. This results in improper identification of boundary pixels in low resolution or blurred images. To reduce these effects, one can adhere to the concept of incorporating local statistics based kernel function in MRF model as $\bar{U}(x_t) = \sum_{\eta_s} e^{-(x_t \otimes h)}$, where η_s denotes a predefined neighborhood of s. This defines the energy function in MRF as a function of x_t convolved with some local statistic-based edge sensitive kernel h [18]. For defining the local statistics of the image, one can consider a Laplacian or Sobel-type edge kernel as in [21]. In this regard, we found that Laplacian edge kernel gives noisy boundary. Similarly, a thicker edge is obtained by Sobel-type edge kernel.

A video scene is likely to contain a number of regions having fairly distinct gray levels as the object is moving from one place to another of a given video. Generally, the change in gray level between the successive regions in an image frame is very common and edge detection techniques are found to be effective only for images with significant contrast. Again, a color image frame possesses high ambiguity within the pixels due to its possible multivalued levels of brightness. Hence, the inclusion of a deterministic edge kernel in MRF modeling is not expected to yield a satisfactory solution. It justifies to apply the concept of fuzzy set [22] based edge kernel [23] rather than a deterministic edge kernel. In this paper, we propose to use fuzzy set-based edge kernel in the MRF modeling.

Use of fuzzy edge kernel at a particular site (with a set of neighboring pixels) makes the energy function look like

$$\bar{U}(x_{st}) = \sum_{\eta_s} e^{-\frac{(x_{st} \otimes h)}{F_c}} \tag{1}$$

where

$$(x_{st} \otimes h) = \frac{1}{\left(1 + \frac{|x_{st} - \bar{x}_{rt}|}{F_d}\right)^{F_e}}. \tag{2}$$

Here, F_e and F_d are two positive constants and are termed the exponential and denominational fuzzifiers, respectively. The constant F_c is an MRF convergence parameter. The term $\bar{x}_{rt}$ of (2) is represented as

$$\bar{x}_{rt} = \max_{r \epsilon \eta_s}\{x_{rt}\} \quad \text{or} \quad \min_{r \epsilon \eta_s}\{x_{rt}\}.$$

Hence, we may write the prior probability $P(X_t)$ as

$$P(X_t) = e^{-\bar{U}(x_t)} = \frac{1}{z} \sum_{\eta_s} e^{-\frac{(x_t \otimes h)}{F_c}}. \tag{3}$$

In the considered expression (3), there are three constants F_c, F_e, and F_d. There is no closed-form solution for estimating the parameters F_e, F_d, and F_c. Hence, in the proposed scheme we have manually fixed these parameters on a trial and error basis.

The expression for MRF modeling with deterministic edge kernel can be given by

$$P(X_t) = e^{-\bar{U}(x_t)} = \frac{1}{z} \sum_{\eta_s} e^{-\frac{\nabla x_{st}}{F_c}} \tag{4}$$

where ∇x_{st} represents the gradient edge computed over a particular site s of the image frame x_t.

A standard problem in MRF-based segmentation is determination of MRF model bonding parameter. The use of a deterministic edge kernel for penalizing the MRF model bonding parameter (α) always produces biased segmentation results. It gets worsen if the images have distinctive peaks in the histograms, and hence make the resulting α value unfit for the model. The obscured or blurred boundary contains gray level ambiguity. It is also true that the use of deterministic edge kernel rarely gives satisfactory results in detecting obscured edges from noisy and blurred scenes. Use of fuzzy sets [23], [24] is found to provide satisfactory results in this regard. Fuzzy set theories are reputed to handle uncertainties to a reasonable extent, arising from deficiencies of information available from a situation (the deficiency may result from incomplete, ill-defined, not fully reliable, vague and contradictory information) [24]. The utility of π function [as in (2)] for modeling obscured edges of noisy images is already established [23]. Hence, incorporation of fuzzy models will have less chance of over-segmentation in spatial segmentation and is expected to provides less object background misclassification.

C. MAP Estimation Framework

Here, the 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, we require to estimate $\hat{x}_t$ from y_t. One way to estimate $\hat{x}_t$ is based on the statistical MAP estimation 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

$$\hat{x}_t = \arg \max_{x_t} P(X_t = x_t | Y_t = y_t). \tag{5}$$

Since x_t is unknown, it is difficult to evaluate (5). Using Bayes' theorem, (5) can be written as

$$\hat{x}_t = \arg \max_{x_t} \frac{P(Y_t = y_t | X_t = x_t) P(X_t = x_t)}{P(Y_t = y_t)}. \tag{6}$$

Since y_t is known, the marginal probability $P(Y_t = y_t)$ is constant. Hence, (6) reduces to

$$\hat{x}_t = \arg \max_{x_t} P(Y_t = y_t | X_t = x_t, \theta) P(X_t = x_t) \tag{7}$$

where θ is the parameter vector associated with estimation of x_t. In (7), $P(X_t = x_t)$ is the prior probability and $P(Y_t = y_t|X_t = x_t, \theta)$ is the likelihood function.

The likelihood function $P(Y_t = y_t|X_t = x_t, \theta)$ can be expressed as

$$P(Y_t = y_t|X_t = x_t, \theta) = P(Y_t = x_t + n|X_t = x_t, \theta) \\ = P(N = y_t - x_t|X_t = x_t, \theta)$$

where n is a realization of Gaussian noise $N(\mu, \Sigma)$. Thus, $P(Y_t = y_t|X_t = x_t)$ can be expressed as

$$P(N = y_t - x_t|X_t, \theta) = \\ \frac{1}{\sqrt{(2\pi)^f \det\left[\Sigma_Q\right]}} e^{-\frac{1}{2}(y_t-\mu_Q)^T \Sigma_Q^{-1}(y_t-\mu_Q)} \tag{8}$$

where Σ is the covariance matrix, det$[\Sigma]$ represents the determinant of matrix Σ, and f is the number of features (for color image, RGB are the three features and it is assumed that there is decorrelation among the three RGB planes). Each pixel (i.e., site s) class in frame t is $Q_{ts} = \{q_1, q_2, ..., q_m\}$ represented by its mean vector $\mu_{Q_{ts}}$ and covariance matrix $\Sigma_{Q_{ts}}$. The likelihood function $P(Y_t = y_t|X_t = x_t)$ can be expressed as

$$P(Y_t = y_t|X_t = x_t, \theta) = \\ \prod_{s\epsilon S} \frac{1}{\sqrt{(2\pi)^f \det\left[\Sigma_{Q_{ts}}\right]}} e^{-\frac{1}{2}(y_t-\mu_{Q_{ts}})^T \Sigma_{Q_{ts}}^{-1}(y_t-\mu_{Q_{ts}})}. \tag{9}$$

Now, putting (9) and (3) in (7), we get

$$\hat{x}_t = \arg\max_{x_t} \\ \left\{\sum_{s\in S} \frac{1}{\sqrt{(2\pi)^f \det[\sum_{Q_{ts}}]}} e^{-\frac{1}{2}(y_t-\mu_{Q_{ts}})^T \sum_{Q_{ts}}^{-1}(y_t-\mu_{Q_{ts}})} \right. \\ \left. - \sum_{\eta_s} \frac{(x_t \otimes h)}{F_c} \right\}. \tag{10}$$

This can be rewritten as

$$\hat{x}_t = \arg\max_{x_t} \\ \left\{\sum_{s\in S} \left\{A - \left[\tfrac{1}{2}(y_t - \mu_{Q_{ts}})^T \textstyle\sum_{Q_{ts}}^{-1}(y_t - \mu_{Q_{ts}})\right]\right\} \right. \\ \left. - \left[\sum_{\eta_s} \frac{(x_t \otimes h)}{F_c}\right]\right\} \tag{11}$$

where $A = -\frac{1}{2}\log((2\pi)^f \det\left[\Sigma_{Q_{ts}}\right])$. $\hat{x}_t$ in (11) is the MAP estimate of x_t. Maximization of (11) requires $2^{q\times M\times N}$ possible image configurations to be searched, where q represents the number of bits required to represent the segmented image and $M \times N$ is the size of the image. The parameter $\theta = \{\mu_Q, \Sigma_Q\}$ for each region type is estimated recursively with EM algorithm [25]. We have considered a combination of both simulated annealing and iterated conditional mode algorithm [8], [9] for estimating the MAP of each incoming image frame.

D. EM Algorithm

The MRF model parameters are estimated recursively in EM framework. Here, we describe the incomplete data problem and the EM algorithm in a general framework. In the MRF framework, y_t is considered to be the observed data (i.e., the given image frame) and x_t as the corresponding unobserved (original) data to be estimated (segmentation result).

For estimation of x_t, the incomplete data y_t is modeled with MRF. We have considered $\theta = \{\mu_Q, \Sigma_Q\}$ as the parameter vector associated with the estimated random variable. The aim of EM algorithm is to estimate the parameter $\theta = \{\mu_Q, \Sigma_Q\}$ based on the observed data y_t. The algorithm begins with an initial arbitrary value θ^o at zeroth instant of time and at iteration u, image labels are estimated using the parameters θ^u. The EM algorithm is an iterative scheme that follows two steps as expectation step and maximization step.

In the expectation step, the function Q is calculated as

$$Q(\theta|\theta^u) = E\left\{\log P(X_t = x_t, Y_t = y_t|\theta)|Y_t = y_t, \theta^u\right\}. \tag{12}$$

Equivalently, it may take the form

$$Q(\theta|\theta^u) = \\ \sum_{x_t \epsilon X_t} P(X_t = x_t, Y_t = y_t|\theta^u) \log P(X_t = x_t, Y_t = y_t|\theta^u).$$

In maximization step, the current estimate θ^{u+1} can be obtained by maximizing the following:

$$\theta^{u+1} = \arg\max_{\theta}\left\{Q(\theta, \theta^u)\right\}. \tag{13}$$

In our framework, we have used EM algorithm to estimate the parameter $\{\theta = x_t, k\}$. Using the initial label estimate x^0, and the observed degraded image y_t, the model parameter θ^1 is estimated by maximizing $Q(\theta, \theta^0)$ as in [25]. This process is iterated step by step until the parameters converge to optimal values.

IV. Region Matching-Based Motion Estimation

In the proposed framework, for detecting moving objects in different frames of a given video captured from a motion dominant camera, we have proposed a region matching-based motion estimation scheme. Here, we assumed that each moving object in the candidate frame (i.e., the frame in which the object position is already known) contains a few homogenous regions. As the object in the scene moves from one position to another, the gray level contents/distribution of each region corresponding to that object will remain unaltered.

In the initial phase of the proposed object detection framework, the candidate frame and the target frame (i.e., the frame in which the object is required to be detected) both are segmented (spatially) by the proposed fuzzy edge incorporated MRF model-based technique. As the position of the moving object in the candidate frame is known, different gray-level regions corresponding to the moving object parts in the candidate frame can be obtained. The object detection task is accomplished by matching the gray level distributions of these regions corresponding to each moving object in the candidate frame to a region in the target frame.

One can consider the entire image as the search space for matching a region corresponding to the moving object in the target frame. However, the complexity of searching becomes very high. To minimize this effect, one can consider a *brute force* scheme (where object is assumed to be in the neighborhood of the centroid of the object). In real-life videos it may happen that from one frame to another, the varying shape of objects may have a large movement/displacement. Hence, to get an effective object detection technique we have proposed a new kind of measure that will give a rough estimate of maximum possible shift in object position in the target frame using the amount of shift in the centroid of objects from the candidate to the target frame. This information is used to define a search space in the target frame. The regions corresponding to an object location in the candidate frames are matched with a region inside the search space of the target frame by considering a χ^2-test-based histogram matching scheme.

A. Calculation of Shift in Object Position

The centroid shifting framework is designed to find the maximum possible movements of a region corresponding to moving objects in the target frame. Let us consider two image frames at tth and $(t + d)$th instant of times, where the tth frame is the candidate frame and the $(t + d)$th frame is the target frame. Our aim is to find a rough estimate/idea of the maximum possible shift in object position from tth to $(t+d)$th frame.

Let us consider $(\hat{i}_t, \hat{j}_t)$ to be the centroid of the candidate frame that can be obtained as follows:

$$\hat{i}_t = \frac{\sum_m \sum_n m * y_t(m,n)}{\sum_m \sum_n y_t(m,n)} \quad \hat{j}_t = \frac{\sum_m \sum_n n * y_t(m,n)}{\sum_m \sum_n y_t(m,n)}$$

where (m, n) is the coordinate of a pixel in the image frame. There are d number of moving objects in the scene, and the positions of these d numbers of moving object locations in the candidate frame are known. We can obtain the centroid of each moving object in the scene by darkening the other objects in the candidate frame object location as follows:

$$\hat{i}_{tb} = \frac{\sum_m \sum_n m * c_{tb}(m,n)}{\sum_m \sum_n c_{tb}(m,n)} \quad \hat{j}_{tb} = \frac{\sum_m \sum_n n * c_{tb}(m,n)}{\sum_m \sum_n c_{tb}(m,n)}$$

where $b = 1, 2, ..., d$. The value of $c_{tb}(m, n)$ in the tth frame for the bth object can be obtained as follows:

$$c_{tb}(m,n) = \begin{cases} 1, & \text{if pixel at } (m,n) \text{ is in the } b\text{th object region} \\ 0, & \text{otherwise.} \end{cases}$$

We have computed the distance of each object centroid from the candidate frame centroid as

$$Dist_{t,tb}(i) = |\hat{i}_t - \hat{i}_{tb}| \quad Dist_{t,tb}(j) = |\hat{j}_t - \hat{j}_{tb}|.$$

For the $(t + d)$th frame (target frame), we have computed the centroid as follows:

$$\hat{i}_{(t+d)} = \frac{\sum_m \sum_n m * y_{(t+d)}(m,n)}{\sum_m \sum_n y_{(t+d)}(m,n)}$$
$$\hat{j}_{(t+d)} = \frac{\sum_m \sum_n n * y_{(t+d)}(m,n)}{\sum_m \sum_n y_{(t+d)}(m,n)}.$$

Similarly, the shift in centroid of target frame from the object location in the candidate frame can be obtained as

$$Dist_{(t+d),b}(i) = |\hat{i}_{(t+d)} - \hat{i}_{tb}|, \quad Dist_{(t+d),b}(j) = |\hat{j}_{(t+d)} - \hat{j}_{tb}|.$$

In the $(t + d)$th frame, the maximum possible shift in object in x-direction can be given as

$$shift_{(x,(t+d)b)} = const * max\{|Dist_{(t+d),b}(i) - Dist_{t,b}(i)|_{b=1,2...d}\}. \tag{14}$$

Similarly, in the y-direction it can be given as

$$shift_{(y,(t+d)b)} = const * max\{|Dist_{(t+d),b}(j) - Dist_{t,b}(j)|_{b=1,2...d}\}. \tag{15}$$

Here, the *const* represents a constant that resolves the camera movement and scaling of the object in the scene, and is a $+ve$ constant. For the considered video sequences we have fixed the range of *const* as $1 < const < 5$.

To find the moving object locations in the target frame initially, the centroid corresponding to each object in the candidate frame is located in the target frame. Search will be made up to $shift_{(x,(t+d)b)}$ in the $+ve$ and $-ve$ directions of the x-axis and $shift_{(y,(t+d)b)}$ in the $+ve$ and $-ve$ directions of the y-axis. A region is searched in the target frame where it exactly matches a region corresponding to the object "b" in the candidate frame. It is very difficult to detect moving objects from low illumination videos or objects scaled from one frame to another. Hence, in the proposed scheme, region matching is performed by χ^2-test-based local histogram comparison [26], as described below.

B. Local Histogram Matching by χ^2-Test

For matching each region corresponding to a moving object in the candidate frame to a region within the search space in the target frame, we have used a histogram matching scheme. Here, the region information is obtained by the proposed MRF-fuzzy edge map-based spatial segmentation scheme. Each time, we search for a pixel within the search space in the target frame and test whether the pixel belongs to a region type in the obtained spatial segmentation or not. The histogram corresponding to the new region is then matched with a region corresponding to the object in the candidate frame by χ^2-test [26].

The χ^2-test is a statistical method of determining the similarity between two distributions. By χ^2-test the distribution/histogram of two regions can be compared by

$$\chi^2((h^{(b_g,y_t)}), (h^{(g,y_{(t+d)})})) = \sum_{l=0}^{l=M-1} \frac{|(h^{(b_g,y_t)}) - (h^{(g,y_{(t+d)})})|}{(h^{(b_g,y_t)}) + (h^{(g,y_{(t+d)})})}.$$

The term $h^{(b_g,y_t)}$ represents the histogram corresponding to a region in bth object of the candidate frame (at tth instant of time). The term $h^{(g,y_{(t+d)})}$ represents the histogram corresponding to a region within the search space (as obtained by the centroid shifting scheme) in the target frame (at $(t+d)$th instant of time). Here, g represents any region in the target frame within the obtained search space and $g = 1, 2, ..., p$, where p is the total number of regions. Number of possible gray level bins in the histogram is considered as M. Two regions are said to be similar if $\chi^2((h^{(b_g,y_t)}), (h^{(g,y_{(t+d)})})) \leq \epsilon$, where ϵ is a small $+ve$ constant. The locations of moving objects obtained by the proposed region-based motion estimation scheme are represented by a binary map, where object regions are given white color and background regions are given black color and this map is termed *foreground region map*.

The *foreground region map* corresponding to the moving objects in the target frame is marked with white color. Hence, the obtained *foreground region map* is a binary output with moving objects as one class (denoted by FM_t) and the background as the other class (denoted as BM_t). The regions forming the foreground part in the *foreground region map* are identified as moving object region, and the pixels corresponding to the FM_t part of the original frame y_t form the VOP.

V. Results and Discussion

In this section, we show some experimental results of the proposed algorithm. The algorithm is implemented in C/C++ and is run on Pentium D, 2.8 GHz PC with 2 G RAM and `Ubuntu` operating system. The proposed scheme is tested on several video sequences; however, for space constraints we have provided results on two test video (color) sequences. We have considered one *Wall Flower* video sequence and one UCF sports action sequence [27]. The video sequences are *Diving* and *LightSwitch*. To make the problem more challenging, few frames of these sequences are blurred with a 5×5 averaging filter.

The result analysis section is divided into three parts. In the first part, we have provided results obtained by the proposed fuzzy edge incorporated MRF modeling for spatial segmentation of the considered video frames. To validate the proposed scheme, results obtained by this are compared with those of spatial segmentation by the meanshift scheme [15], the conventional MRF model [13], and the deterministic edge incorporated MRF model-based schemes. In the considered expression (3), we have replaced the fuzzy edge kernel with a deterministic kernel and termed it spatial segmentation by MRF modeling with deterministic edge kernel [as expressed in (4) of Section III-B] and also used it for comparison.

In the second part of the experiment, the region-based motion estimation scheme is used for identification of the moving objects in the considered video frames. The proposed scheme is validated by comparing the results obtained by the proposed scheme with object detection by the optical flow scheme [5], the MRF-based label fusion scheme [13], and the level-set-based scheme [16].

In the third part of the experiment, validation of the proposed scheme is carried out by two performance evaluations, one for spatial segmentation and the other for object detection. The manual constructed ground truth images are considered for evaluating the performance of both the schemes.

A. Visual Analysis of Results

The first video considered for our experiments is *Diving* video sequence having a single moving object, i.e., a "gymnast" diving from the spring board. Here, the moving object is the "gymnast." The considered frames of the sequence are 1st, 7th, 13th, and 19th, as shown in Fig. 2(a). The spatial segmentation results obtained by the meanshift, the conventional MRF model, and MRF model with deterministic edge kernel are shown in Fig. 2(b)–(d), respectively. It is observed from these results that the output obtained by these approaches is over-segmented. This is due to the fact that the gray-level variation of one region to another of the image frames is very less. The boundary of the object in the scene is blurred and hence is indistinguishable. The deterministic edge incorporated MRF scheme is also unable to identify the object boundary accurately. The results obtained by the proposed fuzzy edge incorporated MRF modeling scheme are shown in Fig. 2(e), where the different parts of the image frames were properly segmented as compared to the results obtained by the meanshift, the conventional MRF model [13], and the deterministic edge incorporated MRF model.

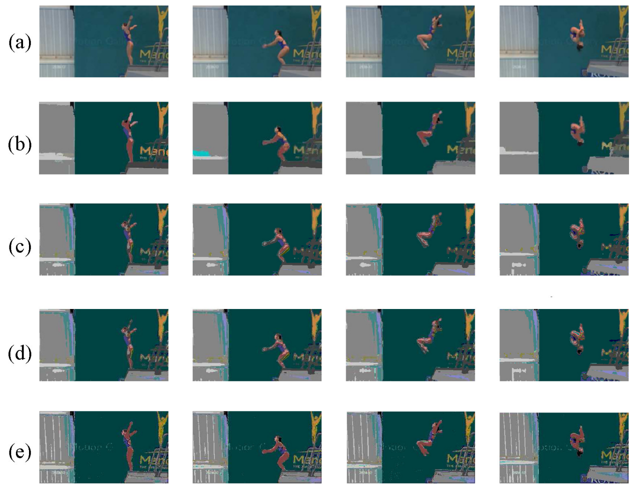


Fig. 2. Spatial segmentation of frame numbers 1st, 7th, 13th, and 19th of the *Diving* video sequence. (a) Original frames. (b) Spatial segmentations using meanshift scheme. (c) Spatial segmentations using only MRF model. (d) Spatial segmentations using MRF and deterministic kernel based local edge prior. (e) Spatial segmentations using proposed fuzzy kernel based local edge prior.

The moving object VOPs obtained for different frames of this sequence using optical flow-based scheme are shown in Fig. 3(b). It is found from these results that the diving gymnast was not properly identified by the optical flow scheme. Similarly, the moving object detection results using the label fusion scheme are shown in Fig. 3(c). It is observed from these results that this technique also was not able to identify different parts of the gymnast correctly. The moving object detection results using the level-set-based scheme are shown in Fig. 3(d), where the gymnast was not properly detected in many frames. However, the results obtained by the proposed scheme [as shown in Fig. 3(e)] have correctly detected the gymnast in different frames of the considered sequence.

The second example we have considered in our experiment is *LightSwitch* sequence having one moving object i.e., a person as shown in frames 1857th, 1858th, 1859th, and 1860th of Fig. 4(a). This sequence is available in [28]. In this video sequence illumination of the scene is changing as the door of the room is opening and closing. We have considered few frames of the video where the person is entering into the room and then the door of the room is getting closed; hence, the reflectance value of each pixel in the scene is

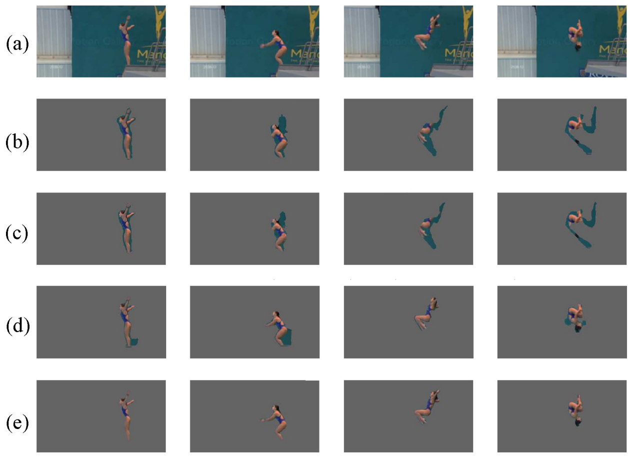


Fig. 3. VOPs of frame numbers 1st, 7th, 13th, and 19th of the *Diving* video sequence. (a) Original frames. (b) VOP generated using the optical flow scheme. (c) VOP generated using the label fusion scheme. (d) VOP generated using the level-set-based method. (e) VOP generated using the proposed scheme.

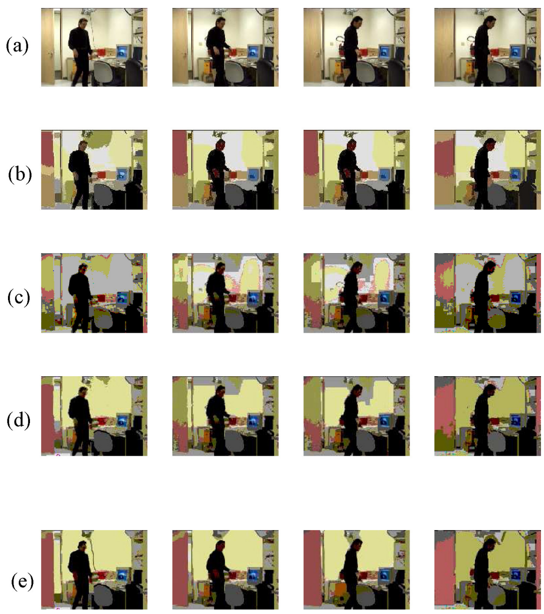


Fig. 4. Spatial segmentation of frame numbers 1857th, 1858th, 1859th, and 1860th of the *LightSwitch* video sequence. (a) Original frames. (b) Spatial segmentations using meanshift scheme. (c) Spatial segmentations using only MRF model. (d) Spatial segmentations using MRF and deterministic kernel-based local edge prior. (e) Spatial segmentations using proposed fuzzy kernel-based local edge prior.

changed accordingly. Spatial segmentation of these frames using meanshift, MRF modeling [13], and deterministic edge-based MRF scheme is shown in Fig. 4(b)–(d), respectively. It is observed from these results that the output obtained by these approaches is over-segmented. The results obtained by the proposed model are shown in Fig. 4(e), where the accuracy of segmentation is more as compared to the results obtained by the meanshift, conventional MRF model [13], and deterministic edge incorporated MRF model.

The detected moving object VOPs obtained for different frames of this sequence using optical flow-based scheme are shown in Fig. 5(b). It is found from these results that the object "man" was not properly identified by the optical flow scheme. Similarly, the moving object detection results using the label fusion scheme are shown in Fig. 5(c). It is observed from these results that this technique also was not able to identify different parts of the person correctly. The results obtained by level-set-based scheme [as shown in Fig. 5(d)] have falsely detected many parts as the object. However, the results obtained by the proposed scheme [as shown in Fig. 5(e)] have detected the moving person in a better way in different frames of the considered sequence.

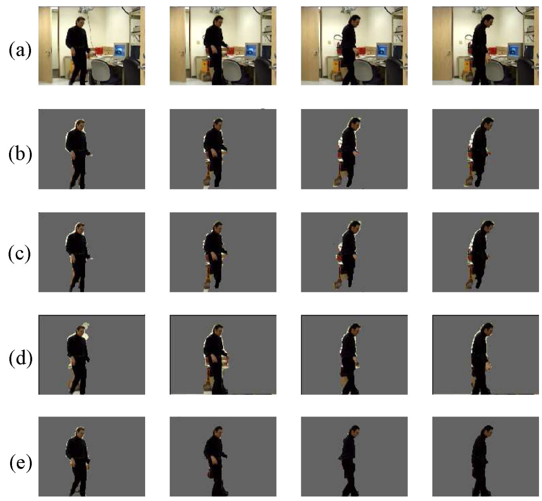


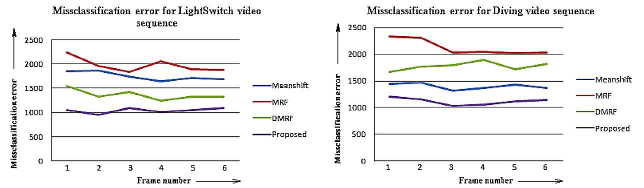
Fig. 5. VOPs of frame numbers 1857th, 1858th, 1859th, and 1860th of the *LightSwitch* video sequence. (a) Original frames. (b) VOP generated using the optical flow scheme. (c) VOP generated using the label fusion scheme. (d) VOP generated using the level-set-based method. (e) VOP generated using the proposed scheme.

B. Quantitative Analysis of Results

Although humans are the best evaluator of any vision system, it is not possible for the human being to evaluate the performance in a quantitative manner. Hence, it is necessary to evaluate methods in an objective way. 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 the other measure is used for quantitative evaluation of the obtained moving object locations. For both these measures we have initially built some manually segmented *ground-truth* images. The results obtained by the proposed scheme and the label fusion scheme are compared with the corresponding *ground-truth* images. For evaluating the accuracy of the proposed spatial segmentation, we have used the pixel-by-pixel comparison of the *ground-truth* images with the obtained spatial segmentation results. This measure is also called *number of misclassified pixels*.

To prove the effectiveness of the proposed scheme, we have plotted the misclassification error for six sampled frames as shown in Fig. 6. It is found that the results obtained by the proposed scheme have higher accuracy.

To evaluate the performance of detected moving object locations, we have considered the performance evaluation



(a) (b)

Fig. 6. Misclassification error curve. (a) *LightSwitch* video sequence. (b) *Diving* video sequence.

TABLE I

PERFORMANCE MEASURE

Video	*PM**	*OF**	*LF**	*LS**	Proposed
LightSwitch	MiP*	0.47	0.37	0.37	**0.21**
	MP*	0.39	0.28	0.18	**0.24**
	SP*	0.16	0.11	0.13	**0.10**
	CP*	0.34	0.26	0.26	**0.18**
Dive	MiP*	0.55	0.51	0.43	**0.42**
	MP*	0.40	0.38	0.28	**0.22**
	SP*	0.40	0.31	0.26	**0.22**
	CP*	0.45	0.39	0.32	**0.29**

*PM: performance measure, OF: optical flow, LF: label fusion, LS: level set, MiP: misclassification penalty, MP: motion penalty, SP: shape penalty, CP: combined penalty.

scheme of Erdem and Sankur [29]. This scheme utilizes four different measures: *misclassification penalty*, *motion penalty*, *shape penalty*, and *combined penalty*. It may be noted that all these measures should be low for proper detection of moving objects. The performance measures for all the video sequences are given in Table I. The performance measures presented in this table are obtained by taking an average of six individual frame's statistics. It is found from this table that the proposed scheme provides less amount of *misclassification penalty*, *motion penalty*, *shape penalty*, and *combined penalty* as compared to the optical flow [5], label fusion [13], and level-set [16] based schemes.

Times required by different techniques, to detect the moving objects per frame from the considered video sequences, are provided in Table II. It is observed from this table that optical flow-based scheme takes very less time for moving object detection. The proposed scheme takes comparable amount of time with that of label fusion scheme and much less time than the level-set-based scheme. This makes the proposed algorithm useful for specific offline applications, such as visual scene analysis, event analysis in surveillance, video annotation, and video motion capture.

We have successfully tested the proposed scheme over 25 video sequences (15 benchmark and 10 real life). We also have tested the proposed fuzzy edge incorporated MRF model for spatial segmentation with the area and variance-based correlation measure [7], [30] for object detection. From the experiments we also found that variance-based measure is quite sensitive to illumination variation and noise. If the object in a video sequence is scaled from one frame to another, the area-based correlation also fails to give satisfactory results. However, the proposed fuzzy edge kernel incorporated MRF model for spatial segmentation with χ^2-test-based local histogram

TABLE II

TIME (IN SECOND) REQUIRED FOR EXECUTION OF DIFFERENT ALGORITHMS PER FRAME

Video	*FN**	*OF**	*LF**	*LS**	Proposed
LightSwitch	1857	6	18	40	23
	1858	6	18	40	23
	1859	6	18	40	23
	1860	6	18	40	23
Diving	1	6	18	60	23
	7	6	18	60	23
	13	6	18	60	23
	19	6	18	60	23

*FN: frame number, OF: optical flow, LF: label fusion, LS: level set.

comparison gives better result with affordable computational complexity.

VI. CONCLUSION

A new region matching-based motion estimation scheme to detect moving objects from given video sequences captured by moving camera was developed. We proposed a spatial segmentation technique by incorporating the fuzzy edge map of pixels in MRF modeling to preserve accurate object boundaries. In this approach, initially, a rough estimate of object locations in the scene is obtained by estimating the amount of shift in the centroid from one frame to another. The location of moving objects in a scene was obtained by χ^2-test-based local histogram matching, which helps to detect objects from a low illumination environment or objects scaled from one frame to another. From the experimental results, we can conclude that the proposed scheme can be successfully applied for detecting moving objects from videos captured by moving camera in low illumination conditions. The spatial segmentation results obtained by the proposed scheme were compared with those of meanshift, conventional MRF modeling and MRF modeling with deterministic edge kernel are found to be better. Similarly, the moving object detection results obtained by the proposed scheme were compared with those obtained by optical flow, label fusion, and level-set schemes, and was found to provide less object-background misclassification. It was also observed that the proposed scheme took more time than optical flow scheme, comparable amount of time with that of label fusion scheme, and less time than the level-set-based scheme.

The proposed scheme does not yield good results, if the object of interest has *cast shadows* in the scene. It also does not provide good results for objects with oclussion/disocclusion. In our future work we will try to solve these problems.

ACKNOWLEDGMENT

The authors appreciate the thorough and constructive comments provided by the reviewers and the editors on this paper. They would also like to thank S. Das for his help during the coding process.

REFERENCES

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

[2] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 611–622, May 2011.

[3] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1128–1139, Aug. 2008.

[4] J. Satake and J. Miura, "Robust stereo-based person detection and tracking for a person following robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 1–6.

[5] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Imaging Understanding Workshop*, 1981, pp. 121–130.

[6] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 610–618, May 1998.

[7] 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.

[8] B. N. Subudhi, P. K. Nanda, and A. Ghosh, "A change information based fast algorithm for video object detection and tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 993–1004, Jul. 2011.

[9] B. N. Subudhi, P. K. Nanda, and A. Ghosh, "Entropy based region selection for moving object detection," *Pattern Recognit. Lett.*, vol. 32, no. 15, pp. 2097–2108, 2011.

[10] C. M. Kuo, C. H. Hsieh, and Y. R. Huang, "Automatic extraction of moving objects for head-shoulder video sequence," *J. Vis. Commun. Image Representation*, vol. 16, no. 1, pp. 68–92, 2005.

[11] D.-S. Lee, S. Yeom, J.-Y. Son, and S.-H. Kim, "Automatic image segmentation for concealed object detection using the expectation-maximization algorithm," *Opt. Express*, vol. 18, no. 10, pp. 10659–10667, 2010.

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

[13] P. M. Jodoin, M. Mignotte, and C. Rosenberger, "Segmentation framework based on label field fusion," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2535–2550, Oct. 2007.

[14] Y. Wang, "Joint random field model for all-weather moving vehicle detection," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2491–2501, Sep. 2010.

[15] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[16] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 3, pp. 266–280, Mar. 2000.

[17] C. E. Erdem, B. Sankur, and A. M. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 937–951, Jul. 2004.

[18] W. Synder, Y. S. Han, G. Bilbro, R. Whitaker, and S. Pizer, "Image relaxation: Restoration and feature extraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 6, pp. 620–624, Jun. 1995.

[19] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Tokyo, Japan: Springer, 2001.

[20] R. B. Potts, "Some generalized order-disorder transformations," in *Proc. Cambridge Philos. Soc.*, vol. 48. 1952, pp. 106–109.

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

[22] L. A. Zadeh, "Fuzzy sets," *Inform. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[23] S. K. Pal and R. A. King, "On edge detection of x-ray images using fuzzy sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 1, pp. 69–77, Jan. 1983.

[24] A. Ghosh, N. S. Mishra, and S. Ghosh, "Fuzzy clustering algorithms for unsupervised change detection in remote sensing images," *Inform. Sci.*, vol. 181, no. 4, pp. 699–715, 2011.

[25] Z. Kato and T. Pong, "A Markov random field image segmentation model for color textured images," *Image Vis. Comput.*, vol. 24, no. 10, pp. 1103–1114, 2006.

[26] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. ECCV*, vol. 3021. 2004, pp. 469–481.

[27] *UCF Sports Action Dataset* [Online]. Available: http://server.cs.ucf.edu/vision/data.html

[28] *Wall Flower* [Online]. Available: http://research.microsoft.com/en-us/um/people/jckrumm/WallFlower/TestImages.htm

[29] C. E. Erdem and B. Sankur, "Performance evaluation metrics for object based video segmentation," in *Proc. Eur. Signal Process. Conf.*, 2000, pp. 917–920.

[30] B. G. Kim and D. J. Park, "Unsupervised video object segmentation and tracking based on new edge features," *Pattern Recognit. Lett.*, vol. 25, no. 15, pp. 1731–1742, Nov. 2004.

Ashish Ghosh (M'09) is currently a Professor with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He has published more than 150 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 was selected as an Associate of the Indian Academy of Sciences, Bangalore, India, in 1997. He is a member of the Founding Team that established the National Center for Soft Computing Research at the Indian Statistical Institute, in 2004, with funding from the Department of Science and Technology, Government of India, and is currently the In-Charge of the center. He is a member of editorial boards of various international journals.

Badri Narayan Subudhi (S'07) received the B.E. degree in electronics and telecommunications from the Biju Patnaik University of Technology, Bhubaneshwar, India, in 2004, and the M.Tech. degree in electronics and system communications from the National Institute of Technology, Rourkela, India, in 2009.

He is currently a Research Scholar with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, with financial support from the Council of Scientific and Industrial Research Scheme. He was a Visiting Scientist with the University of Trento, Trento, Italy, from August 2010 to February 2011. His current research interests include video processing, image processing, machine learning, pattern recognition, and remote sensing image analysis.

Susmita Ghosh received the B.Sc. (hons.) degree in physics and the B.Tech. degree in computer science and engineering from the University of Calcutta, Kolkata, India, in 1988 and 1991, respectively, the M.Tech. degree in computer science and engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 1993, and the Ph.D. degree in engineering from Jadavpur University, Kolkata, in 2004.

She has visited various universities and academic institutes for collaborative research and international projects and has delivered lectures in different countries, including Australia, China, Italy, Japan, and the U.S. She has published many research papers in internationally reputed journals and referred conferences, and has edited three books. She has been a Lecturer and a Reader with the Department of Computer Science and Engineering, Jadavpur University, since 1997. She is currently an Associate Professor with the same department. Her current research interests include genetic algorithms, neural networks, image processing, pattern recognition, soft computing, and remotely-sensed image analysis.