# Compound Markov Random Field Model Based Video Segmentation

Badri Narayan Subudhi and Pradipta Kumar Nanda

**Abstract-- We present a novel approach of video segmentation using the proposed compound Markov Random Field video model. This segmentation scheme is based on the spatio-temporal approach where one MRF model is used to model the spatial image and other two MRF models take care in the temporal directions. In this modeling, edge feature in the temporal direction has been introduced to preserve the edges in the segmented images. The problem is formulated as pixel labeling problem and the pixel labels are estimated using the Maximum a Posteriori (MAP) criterion. The MAP estimates are obtained by the proposed hybrid algorithm. The performance of the proposed method is found to be better than that of JSEG method in terms of percentage of misclassification. Different examples are presented to validate the proposed approach.**

***Index Terms*—Covariance matrices, Feature extraction, Gaussian distribution, Gaussian process, Image edge analysis, Image segmentation, MAP Estimation, Modeling, pattern recognition, Simulated Annealing.**

## I. Introduction

There has been a growing research interest in video image segmentation over the past decade and towards this end, a wide variety of methodologies have been developed [1]-[4]. The video segmentation methodologies have extensively used stochastic image models, particularly Markov Random Field (MRF) model, as the model for video sequences [5]-[7]. MRF model has proved to be an effective stochastic model for image segmentation [8]-[10] because of its attribute to model context dependent entities such as image pixels and correlated features. In Video segmentation, besides spatial modeling and constraints, temporal constraints are also added to devise spatio-temporal image segmentation schemes. An adaptive clustering algorithm has been reported [5] where temporal constraints and temporal local density have been adopted for smooth transition of segmentation from frame to frame. Spatio-temporal segmentation has also been applied to image sequences [11] with different filtering techniques. Extraction of moving object and tracking of the same has been achieved in spatio-temporal framework [12] with Genetic algorithm serving as the optimization tool for image segmentation. Recently, MRF model has been used to model spatial entities in each frame [12] and Distributed Genetic algorithm (DGA) has been used to obtain segmentation. Modified version of DGA has been proposed [6] to obtain segmentation of video sequences in spatio-temporal framework. Besides, video segmentation and foreground subtraction has been achieved using the spatio-temporal notion [13]-[14] where the spatial model is the Gibbs Markov Random Field and the temporal changes are modeled by mixture of Gaussian distributions. Very recently, automatic segmentation algorithm of foreground objects in video sequence segmentation has been proposed [15]. In this approach, first region based motion segmentation algorithm is proposed and thereafter the labels of the pixels are estimated.

In this paper, we propose a compound MRF model for segmentation of video sequence in spatio-temporal framework. The spatial entities in a given image frame is modeled as MRF model. Line fields are incorporated to preserve the edges. The temporal direction attributes are incorporated by adhering to another MRF model in the temporal directions. In order to improve the quality of segmentation, edge features in the temporal directions have been incorporated and another MRF model is used to model these edge features. There are three MRF models taking care of spatio-temporal modeling and incorporating an edge feature in the temporal direction to enhance the segmentation accuracy. Thus, a compound MRF model has been used to model the image sequences. The segmentation problem is formulated as a pixel labeling problem and the pixel labels estimation problem is cast in MAP framework. The MAP estimates are obtained by minimizing the energy function of the posterior distributions. By and large the Simulated Annealing (SA) algorithm [16] is used to obtain the MAP estimates, instead we have proposed a hybrid algorithm based on local global attributes to obtain the MAP estimates and hence segmentation. The proposed scheme has been tested for a wide verity of sequences and it is observed that with the proposed edge based compound MRF model yields better segmentation results than that of edgeless model. The ground truth image is constructed manually and the percentage of misclassification is obtained based on the ground truth images. The proposed method is compared with

B. N. Subudhi is with Image Processing and Computer Vision Lab, Department of Electrical Engineering, National Institute of Technology, Rourkela, Orissa, India. (e-mail: subudhi.badri@gmail.com)

Dr. P. K. Nanda is with Image Processing and Computer Vision Lab, Department of Electronics and Telecommunication Engineering, C. V. Raman College of Engineering, Bhubaneswar, Orissa, India. (e-mail: pknanda_d13@yahoo.co.in)

JSEG [17] method and it is found that the proposed method outperformed JSEG in terms of misclassification error.

## II. SPATIO TEMPORAL IMAGE MODELING

Let the observed video sequences $y$ be considered to be 3-D volume consisting of spatio-temporal image frames. For video, at a given time't' $y_t$ represents the image at time't' and hence $y_t$ is a spatial entity. Each pixel in $y_t$ is a site s denoted by $y_{st}$ and hence, $y_{st}$ refers to a spatio-temporal representation of the 3-D volume video sequences $y$.

Let $x$ denote the segmented video sequences and $x_t$ denote the segmentation of each video frame $y_t$. Instead of modeling the video as a 3-D model we adhere to a spatio-temporal modeling. We model $X_t$ as a Markov random Field Model and the temporal pixels are also modeled as MRF. In particular for second order modeling in the temporal directions, we take $X_t$, $X_{t-1}$ and $X_{t-2}$. In order to preserve the edge features, another MRF model is considered for the pixel of the current frame $x_{st}$ and the line fields of $X_{t-1}$ and $X_{t-2}$. Thus, three MRF models are used as the spatio-temporal image model. The two temporal direction MRF models are shown in Fig. 1. (a) and (b). Fig. 1. (a) correspond to the interaction of pixel $x_{st}$ with the corresponding pixels of $x_{t-1}$ and $x_{t-2}$ respectively. The MRF model taking care of edge features, in other words the line fields of frame $x_{t-1}$ and $x_{t-2}$ together with $x_t$ are modeled as MRF. It is known that if $X_t$ is MRF then, it satisfies the markovianity property in spatial direction

$$P\left(X_{st} = x_{st} \middle| X_{qt} = x_{qt}, \forall\, q \varepsilon S, s \neq q\right)$$
$$= P\left(X_{st} = x_{st} \middle| X_{qt} = x_{qt,}(q,t)\varepsilon\eta_{s,t}\right)$$

Where $\eta_{st}$ is denoted the neighborhood of (s, t) and S denotes spatial Lattice of the frame $X_t$. For temporal MRF, the following Markovianity is satisfied.

$$P\left(X_{st} = x_{st} \middle| X_{pq} = x_{pq}, q \neq t, p \neq s, \forall (s,t)\varepsilon V\right)$$
$$= P\left(X_{st} = x_{st} \middle| X_{pq} = x_{pq,}(p,q)\varepsilon\eta_{s,t}\right)$$

where V denotes the 3-D volume of the video sequence. In spatial domain $X_t$ is modeled as MRF and hence the prior probability $P(X_t)$ can be expressed as Gibb's distributed which can be expressed as

$$P(X_t) = \frac{1}{z} e^{\frac{-U(X_t)}{T}}$$

where z is the partition function which is expressed as $z = \sum_x e^{\frac{-U(x_t)}{T}}$, $U(X_t)$ is the energy function and expressed as $U(x_t) = \sum_{c\varepsilon C} V_c(x_t)$ and $V_c(x_t)$ denotes the clique potential function, T denotes the temperature and is considered to be unity.

We have considered the following clique potential function.

$$V_c(x) = \begin{cases} +\alpha, if\ x_{st} \neq x_{pt}\ and\ (s,t),(p,t)\varepsilon S \\ -\alpha, if\ x_{st} = x_{pt}\ and\ (s,t),(p,t)\varepsilon S \end{cases}$$

Analogously in the temporal direction

$$V_{tec}(x) = \begin{cases} +\beta, if\ x_{st} \neq x_{pt}\ and\ (s,t),(p,t)\varepsilon S \\ -\beta, if\ x_{st} = x_{pt}\ and\ (s,t),(p,t)\varepsilon S \end{cases}$$

$$V_{teec}(x) = \begin{cases} +\gamma, if\ x_{st} \neq x_{pt}\ and\ (s,t),(p,t)\varepsilon S \\ -\gamma, if\ x_{st} = x_{pt}\ and\ (s,t),(p,t)\varepsilon S \end{cases}$$

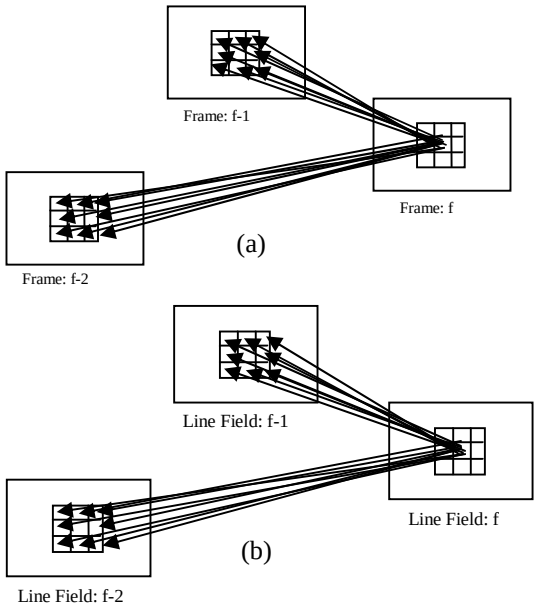



Fig.1. (a) MRF modeling taking two previous frames in the temporal direction (b) MRF with two additional frames with line fields to take care of edge features

### A. Segmentation in MAP framework

The Segmentation problem is cast as a pixel labeling problem. Let $y$ be the observed video sequence and $y_t$ be an image frame at time t and s denote the site of the image $y_t$. Correspondingly $Y_t$ is modeled as a random field and $y_t$ is a realization frame at time t. Thus, $y_{st}$ denotes as a spatio-temporal co-ordinate of the grid (s, t). Let X denotes the segmentation of the video sequence and let $X_t$ denote the

segmentation of an image at time t. Let $X_t$ denote the random field in the spatial domain at time t. The observed image sequences $Y$ are assumed to be the degraded version of the segmented image sequences X. For example at a given time t, the observed frame $Y_t$ is considered as the degraded version of the original label field $X_t$. This degradation process is assumed to be Gaussian Process. Thus, the label field $X_t$ can be estimated from the observed random field $Y_t$. The label field is estimated by maximizing the following posterior distributions.

$$\hat{x} = \arg \max_x P(X = x | Y = y) \quad (1)$$

Where $\hat{x}$ denotes the estimated labels. Since, $x$ is unknown it is very difficult to evaluate (1), hence, using Baye's theorem (1) can be written as

$$\hat{x} = \arg\max_x \frac{P(Y = y | X = x)P(X = x)}{P(Y = y)} \quad (2)$$

Since y is known, the prior probability $P(Y = y)$ is constant. Hence, (2) reduces to

$$\hat{x} = \arg\max_x P(Y = y | X = x, \theta)P(X = x, \theta) \quad (3)$$

Where $\theta$ is the parameter vector associated with $x$.
According to Hammerseley Clifford theorem, the prior probability $P(X = x, \theta)$ is Gibb's distributed and is of the following form

$$P(X = x) = e^{-U(x,\theta)} = e^{\left[-\sum_{c\alpha C}[V_{sc}(x)+V_{tec}(x)+V_{teec}(x)]\right]} \quad (4)$$

In (4) $V_{sc}(x)$ denotes the clique potential function in the spatial domain at time t, $V_{tec}(x)$ denotes the clique potential in the temporal domain and $V_{teec}(x)$ denotes the clique potential in the temporal domain incorporating edge feature. We have proposed this additional feature in the temporal direction. (4) is called the edgebased model. The corresponding edgeless model is

$$P(X = x) = e^{-U(x,\theta)} = e^{\left[-\sum_{c \varepsilon C}[V_{sc}(x)+V_{tec}(x)]\right]}$$

The likelihood function $P(Y = y | X = x)$ can be expressed as

$$P(Y = y | X = x) = P(y = x + n | X = x, \theta) = p(n = y - x | X = x, \theta)$$

Since n is assumed to be Gaussian and there are three components present in color, $P(Y = y | X = x)$ Can be expressed as

$$P(n = y - x | X, \theta) = \frac{1}{\sqrt{(2\pi)^n \det[k]}} e^{-\frac{1}{2}(y-x)^T K^{-1}(y-x)} \quad (5)$$

Where k is the covariance matrix. Assuming decorrelation of the three RGB planes and the variance to be same among each plane, (5) can be expressed as

$$P(n = y - x | X, \theta) = \frac{1}{\sqrt{(2\pi)^3 \sigma^3}} e^{-\frac{1}{2\sigma^2}(x-z)^2} \quad (6)$$

In (6) Variance $\sigma^2$ corresponds to the Gaussian degradation. Hence (3) can be expressed as

$$\hat{x} = \arg\max_x \frac{1}{\sqrt{(2\pi)^3 \sigma^3}} e^{\left(\frac{-\|y-x\|^2}{2\sigma^2}\right)\left(-\left[\sum_{c\varepsilon C} V_{sc}(x)+V_{tec}(x)+V_{teec}(x)\right]\right)} = \arg\max_x e^{-\left[\frac{\|y-x\|^2}{2\sigma^2}+\sum_{c\varepsilon C} V_{sc}(x)+V_{tc}(x)+V_{tec}(x)\right]} \quad (7)$$

Maximizing (7) is tantamount to minimizing the following

$$\hat{x} = \arg\min_x -\left[\frac{\|y-x\|^2}{2\sigma^2}+\sum_{c\varepsilon C} V_{sc}(x)+V_{tc}(x)+V_{tec}(x)\right] \quad (8)$$

$\hat{x}$ in (8) is the MAP estimate and the MAP estimate is obtained by the proposed hybrid algorithm. The associated clique potential parameters and the noise standard deviation $\sigma$ are selected on trial and error basis.

*B. Hybrid Algorithm*

It is observed that SA algorithm takes substantial amount of time for convergence. This algorithm also helps to come out of the local minima and converge to the global optimum solution. This feature could be attributed to the acceptance criterion (acceptance with a probability). We have exploited this feature that is the proposed hybrid algorithm uses the notion of acceptance criterion to come out of the local minima. Subsequently, it is assumed that the solution is locally available and hence a local convergent based strategy is adopted for quick convergence. We have used the Iterated Conditional Mode (ICM) [9] as the locally convergent algorithm. A specific number of iterations is fixed by trial and error. This avoids the undesirable time taken by SA when the solution is close to the optimal solution. The steps of the proposed hybrid algorithm are enumerated as below:

1. Initialize the temperature $T_{in}$.
2. Compute the energy ***U*** of the configuration.
3. Perturb the system slightly with suitable Gaussian

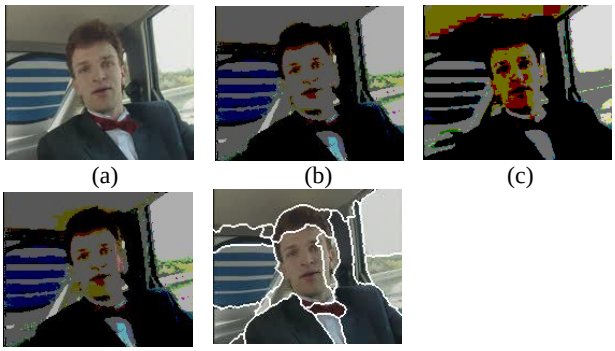
disturbance.

4. Compute the new energy ***U'*** of the perturbed system and evaluate the change in the energy $\Delta U = U' - U$ .
5. If $\Delta U < 0$, accept the perturbed system as the new configuration. Else accept the perturbed system as the new configuration with a probability exp $(-\Delta U / t)$, Where t is the temperature of the cooling schedule.
6. Decrease the temperature according to the cooling schedule**.**
7. Repeat steps 2-7 till some pre specified number of epochs are completed.
8. Compute the energy ***U*** of the configuration.
9. Perturb the system slightly with suitable Gaussian disturbance.
10. Compute the new energy ***U'*** of the perturbed system and evaluate the change in the energy $\Delta U = U' - U$.
11. If $\Delta U < 0$, accept the perturbed system as the new Configuration, otherwise retain the original configuration.
12. Repeat steps 8-12, till the stopping criterion is met. The stopping criteria is the energy U< threshold.

## III. SIMULATION

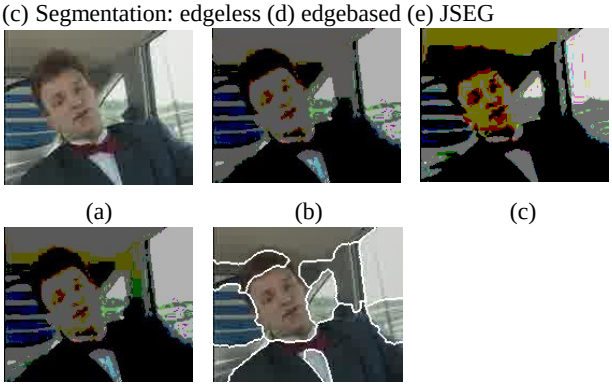
In our simulation, we have considered several video sequences to validate the proposed approach. However for the sake of illustration, segmentation of three video sequences are presented. We have considered video sequence images of spatial size (176x144) and at different times. We have considered two frames in our simulation. Fig. 2 and 3 show the $7^{th}$ and 94th frames and the results obtained. In order to compute the percentage of misclassification error, ground truth images for respective frames have been constructed manually. The edgeless approach and edge based approach is applied and the corresponding segmentation results are shown in Fig. 2. (c) and (d). The model parameters $\alpha$, $\beta$ and $\gamma$ are chosen to be 0.01, 0.009 and 0.007. The standard deviation $\sigma$ of the degradation process is 4.47. The proposed hybrid algorithm has been applied to obtain the MAP estimate of the labels. It is observed from Fig. 2. (c) and (e) that the roof of car and the scene outside the car has been over segmented and hence misclassification of pixels. The proposed edge based model has been used and corresponding segmented results are shown in Fig. 2. (d) for frame No. 7 and Fig. 3. (d) for frame No. 94 respectively. Here, the roof and the scene visible outside this window are classified accurately. The JSEG based result for both the frames are shown in Fig. 2. (e) and 3. (e) and the percentage of misclassification is 4% to 7.54%. It is also observed from Fig. 2. (e) and 3. (e) that whole of face is classified as one class and the car window and the scene outside has been classified as one class and hence there are more misclassified pixels. The MAP estimates are obtained by the proposed hybrid algorithm and Simulated annealing algorithm. As observed from Fig. 8, in case of edge based approach SA converges after 1500 iteration while the hybrid algorithm converges around 300 iteration. In case of edgeless approach, as seen from Fig. 9, hybrid algorithm converges at around 1000 iteration while SA converges at around 2500 iterations. Thus, in both the case hybrid algorithm is faster that of SA.

We have also considered two other examples as shown in Fig. 4 and Fig. 6. The model parameters for this image are $\alpha$, $\beta$ and $\gamma$ are 0.001, 0.008, and 0.006. As observed from Fig. 4. (d) and 5. (d), edgebased model yielded better result than that of edgeless approach. This is also reflected from the percentage of misclassification, given in Table. I. As observed from Table. I, the error is high in case of JSEG method. The third example considered is shown in Fig. 6. The model parameter $\alpha$, $\beta$ and $\gamma$ are 0.009, 0.007 and 0.001. As observed from Fig. 6. (d), the edge based approach could preserve edges and classify better than that of edgeless approach. This phenomenon is also observed in case of 67th frame as shown in Fig. 7. In this case JSEG has also high percentage of misclassification error. Thus, in case of all the examples edge based approach outperformed the edgeless and JSEG method.



(a) (b) (c) (d) (e)

Fig2. Car phone Frame No.7 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG



(a) (b) (c) (d) (e)

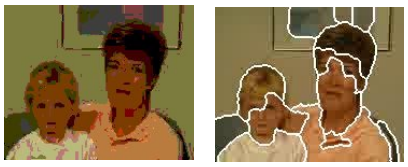
Fig3. Car phone Frame No.94 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG

(a) (b) (c)

(d) (e)

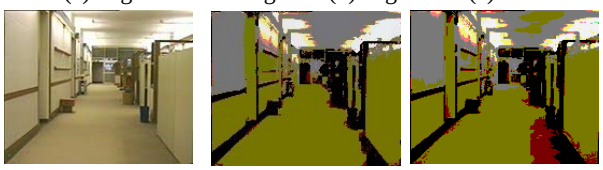
Fig4. Mother Baby Frame No.5 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG

(a) (b) (c)

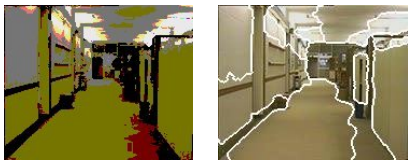


(d) (e)

Fig5. Mother Baby Frame No.65 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG



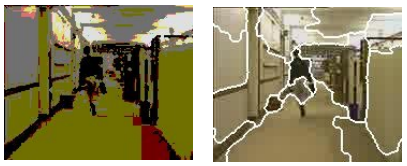
(a) (b) (c)



(d) (e)

Fig6. Hall Monitoring Frame No.6 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG

(a) (b) (c)



(d) (e)

Fig7. Hall Monitoring Frame No.67 (a) original frame (b) Ground Truth (c) Segmentation: edgeless (d) edgebased (e) JSEG

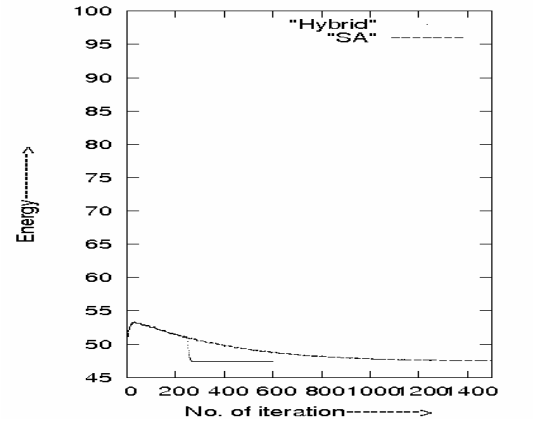



Fig. 8. Graph showing energy convergence for car phone Frame No.7 with Edge Feature and Hybrid Algorithm, with Edge Feature and SA.

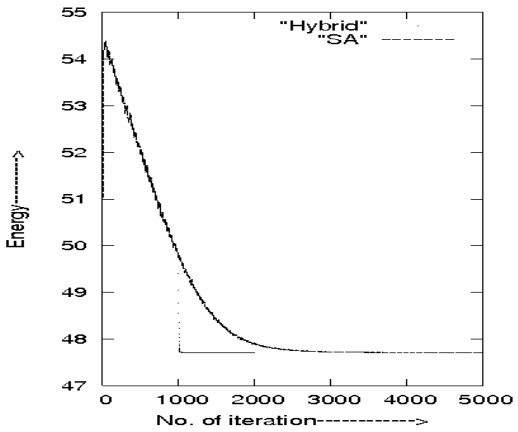



Fig. 9. Graph showing energy convergence for car phone Frame No.7 Edgeless Feature and Hybrid Algorithm, Edgeless Feature and SA

TABLE I
CALCULATION OF MISCLASSIFICATION ERROR

| Video Name | Car Phone Frames | | Mother Baby Frames | | Hall Monitoring | |
|---|---|---|---|---|---|---|
| Frame No | 7 | 94 | 5 | 65 | 6 | 67 |
| Edgeless | 1.3 | 8.03 | 0.27 | 1.1 | 1.0 | 0.92 |
| Edgebase | 0.01 | 2.24 | 0.18 | 0.24 | 0.05 | 0.04 |
| JSEG | 4.0 | 7.54 | 7.54 | 4.7 | 4.55 | 6.75 |

## IV. CONCLUSION

We have proposed a compound MRF model as the spatio-temporal model for video segmentation. This new model takes into account the edge features besides the temporal MRF model. The new model has proved to be an effective model for video segmentation. The problem is formulated as a pixel labeling problem. The pixel labels are estimated using the proposed hybrid algorithm. The hybrid algorithm, exploring local-global feature, is found to converge much faster than that of SA algorithm. The edge based model with hybrid algorithm

is compared with edgeless model and JSEG method and it is found that the edge based model is the best one. The model parameters in all the cases are selected on an adhoc manner. The current work includes the model parameter estimation, motion estimation and tracking of the moving object.

## V. ACKNOWLEDGMENT

The authors acknowledge the facility provided at IPCV Lab of N. I. T, Rourkela and C. V. Raman College of Engineering, Bhubaneswar.

## VI. REFERENCES

[1] A. M. Teklap, *Digital Video Processing*. Prentice Hall, NJ, 1995.

[2] P. Salember and F. Marques, "Region based representation of image and video segmentation tools for multimedia services,"*IEEE Trans. Circuit systems and video Technology*, vol. 9, No. 8, pp. 1147-1169, Dec.1999.

[3] E. Y. Kim, S. H. Park and H. J. Kim, "A Genetic Algorithm-based segmentation of Random Field Modeled images," *IEEE Signal processing letters*, vol. 7, No. 11, pp. 301-303, Nov. 2000.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, No. 6, pp. 721-741, Nov. 1984.

[5] R. O. Hinds and T. N. Pappas, "An Adaptive Clustering algorithm for Segmentation of Video Sequences," *Proc. of International Conference on Acoustics, speech and signal Processing, ICASSP*, vol. 4, pp. 2427-2430, May. 1995.

[6] E. Y. Kim and K. Jung, "Genetic Algorithms for video Segmentation," *Pattern Recognition*, vol. 38, No. 1, pp.59-73, 2005.

[7] E. Y. Kim and S. H. Park, "Automatic video Segmentation using genetic algorithms," *Pattern Recognition Letters*, vol. 27, No. 11, pp. 1252-1265, Aug. 2006.

[8] Stan Z. Li, *Markov field modeling in image analysis*. Springer: Japan, 2001.

[9] J. Besag, "on the statistical analysis of dirty pictures," *Journal of Royal Statistical Society Series B (Methodological)*, vol. 48, No. 3, pp.259-302, 1986.

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

[11] G. K. Wu and T. R. Reed, "Image sequence processing using spatiotemporal segmentation," *IEEE Trans. on circuits and systems for video Technology*, vol. 9, No. 5, pp. 798-807, Aug. 1999.

[12] S. W. Hwang, E. Y. Kim, S. H. Park and H. J. Kim, "Object Extraction and Tracking using Genetic Algorithms," *Proc. of International Conference on Image Processing, Thessaloniki, Greece*, vol.2, pp. 383-386, Oct. 2001.

[13] S. D. Babacan and T. N. Pappas, "Spatiotemporal algorithm for joint video segmentation and foreground detection," *Proc. EUSIPCO*, Florence, Italy, Sep. 2006.

[14] S. D. Babacan and T. N. Pappas, "Spatiotemporal Algorithm for Background Subtraction," *Proc. of IEEE International Conf. on Acoustics, Speech, and Signal Processing*, ICASSP 07, Hawaii, USA, pp. 1065-1068, April 2007.

*[15]* S. S. Huang and L. Fu," Region-level motion-based background modeling and subtraction using MRFs," *IEEE Transactions on image Pocessing*, vol.*16, No. 5, pp.1446-1456, May. 2007.*

[16] S. C. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, No. 4598, pp. 671-680, 1983.

[17] Y. Deng and B. S. Manjunath, "Unsupervised Segmentation of color texture regions in image and video," *IEEE Transaction on pattern Analysis and Machine Intelligence*, vol. 23, No. 8, pp. 800-810, 2001.

## VII. BIOGRAPHIES

**Mr. Badri Narayan Subudhi** after obtaining his B. E degree from Biju Patnaik University of Technology, Orissa in Electronics and Telecommunication Engineering, is continuing his M.Tech (Res) in the Department of Electrical Engineering, N. I. T, Rourkela. Currently he is working in the area of Video Processing and its real time implementation. His research interest is Video Segmentation, Pattern Recognition and Image Processing. He is a Student member of IEEE.

**Dr. Pradipta Kumar Nanda** is currently working as a Professor and Head of the Department of Electronics and Telecommunication Engineering, at C. V. Raman College of Engineering, Bhubaneswar. Prior to this, he had served N. I. T, Rourkela from 1986 to 2007 and was the Professor and Head of the Electrical Engineering Department. He has obtained his PhD from Indian institute of Technology, Bombay in the area of Computer Vision and has guided one PhD and 20 M.Tech students. He has completed different MHRD funded project at N. I. T, Rourkela. Currently he is supervising two PhD students and one M.Tech student. He is a fellow IETE. His research interest is Image Processing and Analysis, Computer Vision, Soft computing, Signal Processing.