Object Tracking using SIFT and Kalman Filter

Hemalatha B[#1], Grevas Timi L[*2], Krishnaamirthalaxmi V S[#3]

[#]*Velammal Engineering College*
Chennai,India
[1]hemalathaathi@gmail.com
[3]krishnakal31@gmail.com

[*] *Velammal Engineering College*
Chennai,India
[2]grevas.timi@gmail.com

***Abstract*- Object detection in videos involves verifying the presence of an object in image sequences and locating it after recognition. A major challenge to object tracking is the occlusion of the target object by other objects in the scene. In this paper two features in video is used to detect and track objects: Visual features (such as color, texture and shape) and Motion information. The proposed method for object tracking is based on SIFT (Scale-Invariant Feature Transform) and Kalman filter. The SIFT algorithm detects the invariant feature points which are used in identifying the target object in non-occluded environment. If the SIFT algorithm fails to track the object in case of occlusion, Kalman filter which has the ability to predict the target location is used at such instances to determine the location of the target.**

***Keywords-Object Tracking,Occlusion* SIFT, Kalman Filter**

I.INTRODUCTION

Video surveillance is an active research topic in computer vision that tries to detect, recognize and track objects over a sequence of images and it also makes an attempt to understand and describe object behavior by replacing the aging old traditional method of monitoring cameras by human operators. Object detection and tracking are important and challenging tasks in many computer vision applications such as surveillance, vehicle navigation and autonomous robot navigation. Object detection involves locating objects in the frame of a video sequence. Every tracking method requires an object detection mechanism either in every frame or when the object first appears in the video. Object tracking is the process of locating an object or multiple objects over time using a camera. The high powered computers, the availability of high quality and inexpensive video cameras and the increasing need for automated video analysis has generated a great deal of interest in object tracking algorithms.

There are three key steps in video analysis like detection of interested moving objects, tracking of such objects from each and every frame to frame, and analysis of object tracks to recognize their behavior. Therefore, the use of object tracking is pertinent in the tasks of, motion based recognition. Automatic detection, tracking, and counting of a variable number of objects are crucial tasks for a wide range of home, business, and industrial applications such as security, surveillance, management of access points, urban planning, traffic control, etc. However, these applications were not still playing an important part in consumer electronics. The main reason is that they need strong requirements to achieve satisfactory working conditions, specialized and expensive hardware, complex installations and setup procedures with of supervision qualified workers. Some works have focused on developing automatic detection and tracking algorithms that minimizes the necessity of supervision. They typically use a moving object function that evaluates each hypothetical object configuration with the set of available detections without to explicitly compute their data association. Thus, a considerable saving in computational cost is achieved. In addition, the likelihood function has been designed to account for noisy, false and missing detections. The field of machine (computer) vision is concerned with problems that involve interfacing computers with their surrounding environment. One such problem, surveillance, has an objective to monitor a given environment and report the information about the observed activity that is of significant interest. In this respect, video surveillance usually utilizes electro-optical sensors (video cameras) to collect information from the environment.

In a typical surveillance system, these video cameras are mounted in fixed positions or on pan-tilt devices and transmit video streams to a certain location, called monitoring room. Then, the received video streams are monitored on displays and traced by human operators. However, the human operators might face many issues, while they are monitoring these sensors. One problem is due to the fact that the operator must navigate through the cameras, as the suspicious object moves between the limited field of view of cameras and should not miss any other object while taking it. Thus, monitoring becomes more and more challenging, as the number of sensors in such a surveillance network increases. Therefore, surveillance systems must be automated to improve the performance and eliminate such operator errors. Ideally, an automated surveillance system should only require the objectives of an application, in which real time interpretation and robustness is needed. Then, the challenge is to provide robust and real-time performing surveillance systems at an affordable price.

With the decrease in costs of hardware for sensing and computing, and the increase in the processor speeds, surveillance systems have become commercially available, and they are now applied to a number of different applications, such as traffic monitoring, airport and bank security, etc. However, machine vision algorithms (especially for single camera) are still severely affected by many shortcomings, like occlusions, shadows, weather conditions, etc. As these costs decrease almost on a daily basis, multi-camera networks that utilize 3D information are becoming more available. Although, the use of multiple cameras leads to better handling of these problems, compared to a single camera, unfortunately, multi-camera surveillance is still not the ultimate solution yet. There are some challenging problems within the

surveillance algorithms, such as background modeling, feature extraction, tracking, occlusion handling and event recognition. Moreover, machine vision algorithms are still not robust enough to handle fully automated systems and many research studies on such improvements are still being done. This work focuses on developing a framework to detect moving objects and generate reliable tracks from surveillance video.

The problem is most of the existing algorithms works on the gray scale video. But after converting the RGB video frames to gray at the time of conversion, information loss occurs. The main problem comes when background and the foreground both have approximately same gray values. Then it is difficult for the algorithm to find out which pixel is foreground pixel and which one background pixel. Sometimes two different colors such as dark blue and dark violet, color when converted to gray scale, their gray values will come very near to each other, it can't be differentiated that which value comes from dark blue and which comes from dark violet. However, if color images are taken then the background and foreground color can be easily differentiated. So without losing the color information this modified background model will work directly on the color frames of the video.

II. LITERATURE SURVEY

Real world object tracking in video is essentially needed in many applications today for security and surveillance, medical imaging and video editing, military, automation, traffic monitoring and animated games. Capturing the object of interest under motion is a challenging task which is resolved by many methodologies.

Lipton et al. [15] proposed determination of frame difference with the use of pixel-wise differences between two frame images to extract the moving regions. In another work, Stauffer Grimson et al. [16] proposed a Gaussian mixture model based on background model to detect the object and also proposed background subtraction to detect moving regions in an image by taking the difference between current and reference background image in a pixel-by- pixel. Collins et al. [17], developed a hybrid method that combines three-frame differencing with an adaptive background subtraction model for their VSAM (Video Surveillance and Monitoring) paper. Desa Salih et al [18], proposed a combination of background subtraction and frame difference that improved the previous results of background subtraction and frame difference. Sugandi et al.[19], proposed a new technique for object detection employing frame difference onlow resolution image. Satoh et al. [20], proposed a new technique for object tracking employing block matching algorithm based on PISC image. Beymer Konolige et al. [21],1999 proposed in stereo camera based object tracking, use Kalman filter for predicting the objects position and speed in x-2 dimension. Julio Cezar et al. [22] has proposed a background model, and incorporated a novel technique for shadow detection in gray scale video sequences. Rosals Sclaroff et al.,1999 proposed the use of extended Kalman filter to estimate 3D trajectory of an object from 2D motion. In object detection method, many researchers have developed their methods. Liu et al., 2001 proposed background subtraction to detect moving regions in an image by taking the difference between current and reference background image in a pixel-by-pixel. It is extremely sensitive to change in dynamic scenes derived from lighting and extraneous events etc.

Stauffer Grimson in 1997 proposed a Gaussian mixture model based on background model to detect the object. Lipton et al., 1998 proposed frame difference that makes use of pixel-wise differences between two frame images to extract the moving regions. This method is very adaptive to dynamic environments, but generally does a poor job of extracting all the relevant pixels, e.g., there may be holes left inside moving entities. In order to overcome disadvantage of two-frames differencing, in some cases three-frames differencing is used. A hybrid method was developed by Collins et al., 2000 that combines three-frame differencing with an adaptive background subtraction model for their VSAM (Video Surveillance and Monitoring) paper. The hybrid algorithm successfully segments moving regions in video without the defects of temporal differencing and background subtraction. Desa Salih in 2004 proposed a combination of background subtraction and frame difference that improved the previous results of background subtraction and frame difference. Object tracking methodology in his work describes more about the region based tracking. Region-based tracking algorithms track objects according to variations of the image regions corresponding to the moving objects. For these algorithms, the background image is maintained dynamically and motion regions are usually detected by subtracting the background from the current image. Wren et al.,1997 explored the use of small blob features to track a single human in an indoor environment. In their work, a human body is considered as a combination of some blobs respectively representing various body parts such as head, torso and the four limbs. The pixels belonging to the human body are assigned to the different body parts blobs. By tracking each small blob, the moving human is successfully tracked. McKenna et al., 2000 proposed an adaptive background subtraction method in which colour and gradient information are combined to cope with shadows and unreliable colour cues in motion segmentation. Tracking is then performed at three levels of abstraction: regions, people, and groups. Each region has a bounding box and regions can merge and split. A human is composed of one or more regions grouped together under the condition of geometric structure constraints on the human body, and a human group consists of one or more people grouped together. Sugandi et al., 2007, papered object tracking by employing block matching algorithm based on PISC image and Satoh et al., 2001 implemented object identification employing colour and spatial information of the tracked object. Cheng Chen, 2006 proposed a colour and a spatial feature of the object to identify the track object. The spatial feature is extracted from the bounding box of the object. Meanwhile, the colour features extracted is mean and standard value of each object. Czyz et al., 2007 proposed the colour distribution of the object as observation model. The similarity of the objects measurement using Bhattacharya distance corresponds to the high similarity if distance measured is less. To overcome this problem described above, this paper proposes a new technique for object detection employing frame difference on low resolution image.

With this literature survey, it is found that detecting the object from the video sequence and also track the object is a really difficult task. Object tracking can be a time consuming process due to amount of data that is contained in the video. Some of the main challenges are

- Variation in target pose or target deformations
- Variation in illumination
- Partial or full occlusion of the object
- Noise in images
- Complex object shapes
- Occlusion of an object refers to blocking of one object by another object

III. System Architecture

Object of interest in video frames are tracked in spite of occlusion using various processing of frames before the application of SIFT algorithm and Kalman filter. The object detection system consists of several modules like pre-processing, selection of ROI (Region of Interest), SIFT, Kalman as depicted in Fig.1.

The tracking algorithm begins when a user selects the object the object to track. The SIFT features found in the location of the object are stored. In each frame, Kalman filter makes prediction for a possible location of the object. The prediction of Kalman filter is used when the SIFT algorithm fails due to occlusion of the target object. Otherwise the measurement from SIFT is used to update the Kalman filter.

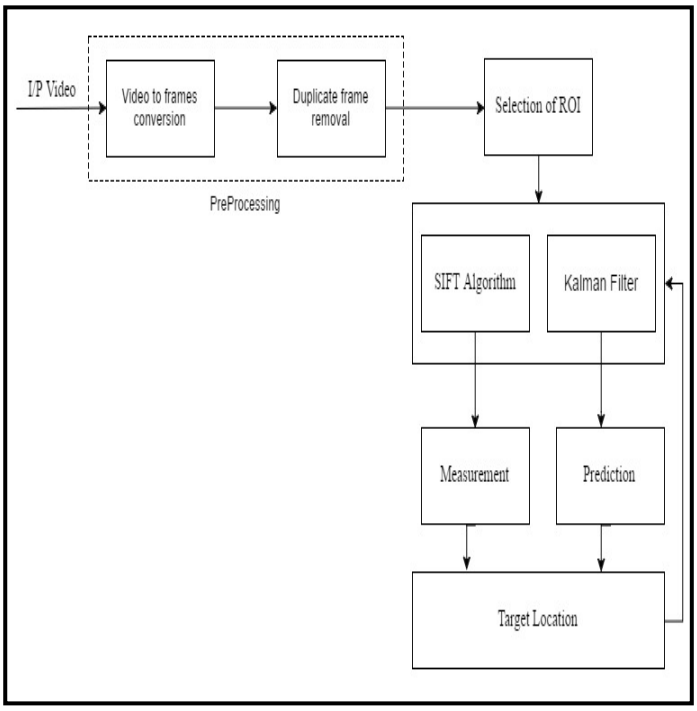

Fig.1.System Architecture

A. Pre-processing

1) Video to frames conversion

Captured video is selected and converted into a sequence of frames for extracting essential features of the object to be tracked. The converted frame from a sample video is shown in Fig. 2.

Fig. 2. Video to duplicate Frames

2) Duplicate frames removal

To avoid redundancy in processing of frames, duplicate frames should be removed from the frame sequence obtained in the previous step. To achieve this, the frames are converted to grayscale and Gaussian blur is applied. Then the absolute difference between the frames is used to identify duplicate frames which are subsequently removed as shown in Fig.3.

Fig.3. Non-duplicate Frames

B. Selection of ROI

The ROI (Region of Interest) is segmented from a frame and the key features for ROI are extracted using SIFT algorithm. The above process for finding key features is applied to other frames. The key features of the frames are matched with ROI key feature using Euclidean distance.

C. SIFT

It is an efficient way to find distinctive local features that are invariant to rotation, scale, and possible occlusion. To find SIFT [1] features, images are produced in different scales. Each image is convolved with a Gaussian kernel, and the differences between adjacent scales of convolved images are calculated. Candidate keypoints are local maxima and minima of the difference. From the candidates, keypoints are selected based on measures of their stability. One or more orientations are assigned to each keypoint location based on local image gradient directions. The gradients at the selected scale in the region will represent the keypoints. This is illustrated in Fig.4.

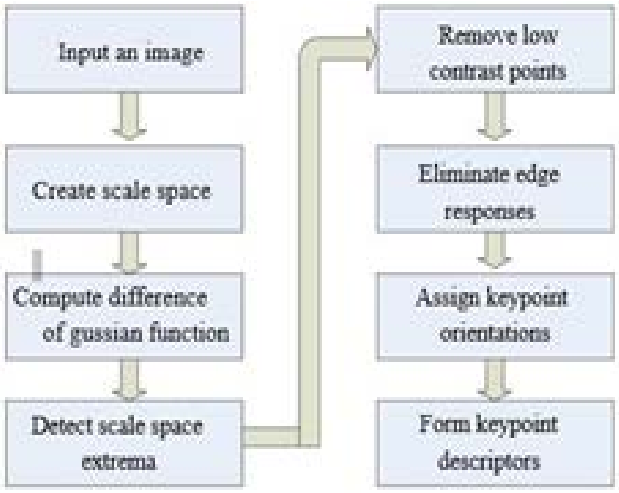

Fig. 4. Processing of Key Points

1) Constructing a scale space

This is the initial preparation. Several octaves of the original image are generated. Each octave's image size is half the previous one. Within an octave, images are progressively blurred using the Gaussian Blur operator. It is suggested that four octaves and five blur levels are ideal for the algorithm. In the next step, these octaves are used to generate Difference of Gaussian images.

2) LoG Approximation

Two consecutive images in an octave are taken and one is subtracted from the other. Then the next consecutive pair is taken, and the process repeats. This is done for all octaves. The resulting images are an approximation of scale invariant Laplacian of Gaussian (which is good for detecting keypoints). These Difference of Gaussian images are approximately equivalent to the Laplacian of Gaussian. These approximations are also scale invariant as in Fig.5.

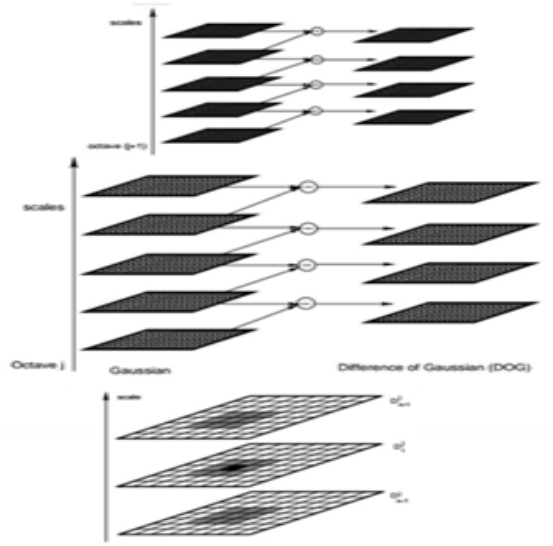

Fig. 5. Scale, octaves and difference of Gaussians

3) Finding Keypoints

Finding key points is a two part process: 1) Locate maxima/minima in DoG images 2) Find subpixel maxima/minima. The first step is to coarsely locate the maxima and minima by iterating through each pixel and checking all its' neighbours. The check is done within the current image, and also the one above and below it. The marked points are the approximate maxima and minima. They are "approximate" because the maxima/minima almost never lies exactly on a pixel. It lies somewhere between the pixel. In the second step, subpixel values are generated by the Taylor expansion of the image around the approximate key point. These subpixel values increase chances of matching and stability of the algorithm.

4) Get rid of bad key points

Edges and low contrast regions are bad keypoints. Eliminating these makes the algorithm efficient and robust. Removing low contrast features: If the magnitude of the intensity at the current pixel in the DoG image (that is being checked for minima/maxima) is less than a certain value, it is rejected. Removing edges: Two gradients are calculated at the keypoint. If both the gradients are big enough (indicating it is corner), the keypoint is passed through, otherwise it is rejected.

5) Assigning an orientation to the keypoints

The gradient directions and magnitudes around each keypoint are calculated around each keypoint. Then the most prominent orientation(s) in that region are figured out. This orientation(s) is assigned to the keypoint. Any later calculations are done relative to this orientation which ensures rotation invariance.

6) Generate SIFT features

A very unique fingerprint is generated for the keypoint sift [3]. To do this, a 16x16 window around the keypoint is considered. This 16x16 window is broken into sixteen 4x4 windows. Within each 4x4 window, gradient magnitudes and orientations are calculated. These orientations are put into an 8 bin histogram. Doing this for all 16 pixels, 128 numbers are generated which are normalized to form the feature vector. The keypoint is uniquely identified by this feature vector.

7) Matching Key Features

Using SIFT feature extraction, the keypoint descriptors for both the target object and the video frames are obtained. For each keypoint descriptor of the target object, the Euclidean distances between that keypoint descriptor and the keypoint descriptors of the video frames are computed. If the distance is less than the threshold (around 0.2-0.7), then it is considered a match sift match[4] and the corresponding match is saved .The matches are plotted using the key point locations. This is repeated for all the video frames as shown if Fig.6. The resulting images are output in the form of a video wherein the target object is identified and marked.

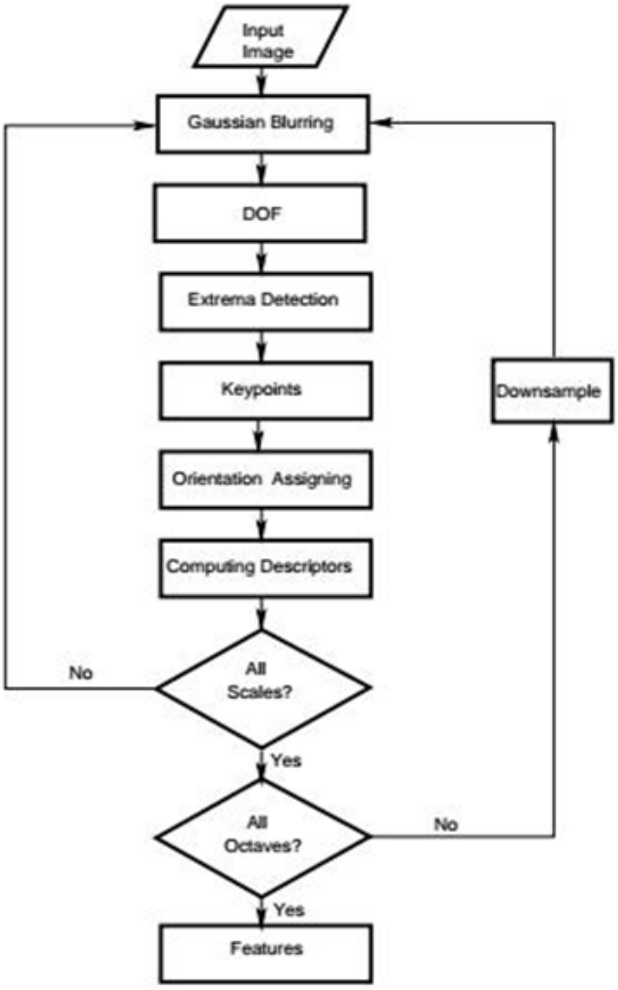

Fig. 6. SIFT Feature Extraction

D. Kalman Filter

Based on the performance analysis of SIFT algorithm, SIFT fails to track the object in case of occlusion. The Kalman filter which has the ability to predict the current state from the motion information of previous state is used for the occluded environment. Hence for videos SIFT algorithm is used for tracking target object based on its invariant features and Kalman Filter is also used for same set of videos for tracking target object using motion information and prediction.

The Kalman filter KF [5] object is designed for object tracking. It used to predict a physical object's future location, to reduce noise in the detected location, or to help associate multiple physical objects with their corresponding tracks. A Kalman filter object can be configured for each physical object for multiple object tracking. To use the Kalman filter, the object must be moving at constant velocity or constant acceleration. The Kalman filter algorithm involves two steps, prediction and correction the first step uses previous states to predict the current state.The second step uses the current measurement, such as object location, to correct the state. The Kalman filter implements a discrete time, linear State-Space System. recursive predictive filter that is based on the use of state space techniques and recursive algorithms

IV. IMPLEMENTATION

The object to be tracked in the video is a ball which is subjected to constant motion and its detection is implemented as below.

The snapshot of the sample video in which the ball is moved is shown in Fig. 7. SIFT algorithm uses the key features extracted from the ball and tracks it as shown by the border around the object in Fig.7.

Fig.7. Tracking of Object

In case of occlusion by covering the ball as shown in Fig.8.,Kalman filter takes over to predict the location of ball based on its previous state.

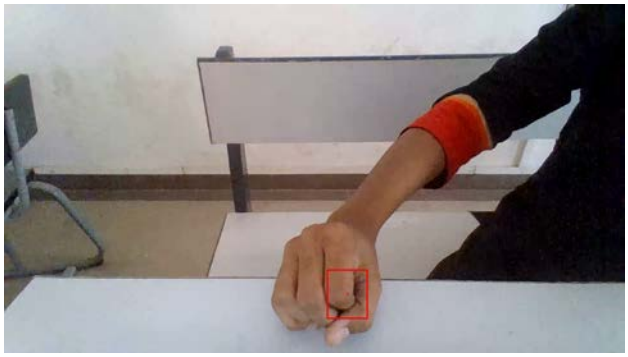

Fig.8. Occlusion of Object

When the image is detected again SIFT again starts to track the object as shown in Fig.9.

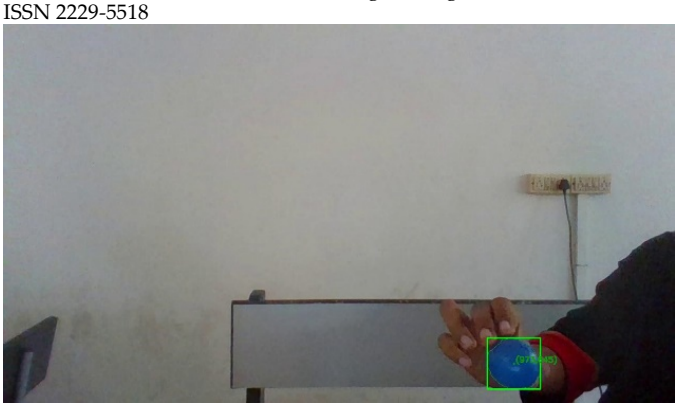

Fig.9. Tracking object after occlusion

V.CONCLUSION

In this work, visual features and motion information of target object is used to track it even under occlusion with the help of SIFT algorithm and Kalman filter. The visual information from frames is used by SIFT while the motion information of object is used by Kalman Filter for object detection. Even if SIFT fails in its attempt to detect object in motion under occlusion Kalman filter

As a part of future work, SIFT and Kalman filter can be combined to work together making Kalman filter to detect object based on SIFT key feature points for improvement in object tracking.

REFERENCES

[1] Song Dan,Zhao Baojun,Tang Linbo ,*A Tracking Algorithm Based on SIFT and Kalman Filter*, The 2nd International Conference on Computer Application and System Modeling (2012).

[2] Seok-Wun Ha, Yong-Ho Moon, *Object Tracking Using SIFT and Kalman Filter* , Department of Informatics, ERI, Gyeongsang National University, Jinju, Rep. of Korea

[3] Afef SALHI and Ameni YENGUI JAMMOUSSI, *Object tracking system using Camshift, Meanshift and Kalman filter*, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering Vol:6, No:4, 2012

[4] Afef SALHI and Ameni YENGUI JAMMOUSSI, Natan Peterfreund, *Robust Tracking of Position and Velocity With Kalman Snakes*.IEEE transactions on pattern analysis and machine intelligence, vol. 21, no. 6, June 1999.

[5] Shoichi Arakil, Takashi Matsuoka, Haruo Takemura, and Naokazu Yokoya,*Real-time Tracking of Multiple Moving Objects in Moving Camera Image Sequences Using Robust Statistics.*1051-4651/98, 1998 IEEE.

[6] R. Curwen and A. Blake,Dynamic Contours*: Real-Time Active Splines.Active Vision*, A. Blake and A. Yuille, eds., pp. 39-58. MIT Press, 1992.

[7] D. Metaxas and D. Terzopoulos,Shape and Nonrigid Motion Estimation Through Physics-Based Synthesis.IEEE Trans. *Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 580-591, 1993.

[8] V. Caselles and B. Coll,Snakes in Movement.SIAM J. *Numerical Analysis*, vol. 33, pp. 2,445-2,456, 1996.

[9] M.Bertalmio, G. Sapiro, and G. Randall, *Morphing Active Contours.Proc. Int'l Conf. Scale-Space Theories in Computer Vision*, pp. 46-57, 1999.

[10] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, *Fast Geodesic Active Contours.Proc. Int'l Conf. Scale-Space Theories in Computer Vision*, pp. 34-45, 1999.

[11] M.P. Dubuisson, S. Lakshmanan, and A.K. Jain, *Vehicle Segmentation and Classification using Deformable Templates.IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 293-308, 1996.

[12] Alan J Lipton, Hironobu Fujiyoshi, and Raju S Patil. *Moving target classification and tracking from real-time video. In Applications of Computer Vision*, 1998. WACV98. Proceedings., Fourth IEEE Workshop on, pages 814. IEEE, 1999.