Moving Object Segmentation in Video Using Stationary Wavelet Transform

Debashis Sen, Ajit Singh Sandhu, Harun Prasad Paramasivam
Department of Electrical and Computer Engineering, Concordia University,
Montreal, Quebec, Canada H3G 1M8.
{d_sen, a_sandhu, h_parama}@ ece.concordia.ca

Abstract

Various video processing applications, such as tracking, requires low complexity and reliable segmentation of objects. Global motion and background clutter often acts as key constraints to perform reliable segmentation. In this paper, we propose a video segmentation algorithm for tracking application that handles these constraints by operating on high and low frequency wavelet bands simultaneously.

Furthermore, our method incorporates novel motion adaptation, clutter removal and region creation techniques. It successfully deals with various types of obstacles, such as large global motion and high background clutter. Simulation results demonstrate that the proposed algorithm achieves appropriate performance in segmentation, at a low complexity level.

1. Introduction

Many object segmentation algorithms incorporating global motion have been proposed [1], [2], [3], [4] and [5], but they are either computationally expensive or suffer from quality problems. The algorithm proposed in [4] uses watershed transformation and fast motion estimation. But, it requires many preprocessing steps, e.g., prefiltering and elimination of small local minima, which is unreliable and computationally expensive. The algorithm in [5] uses a computationally expensive recursive technique to detect multiple motions. In [2], a mixture of computationally expensive K-Gaussian background model is used. The technique in [3] uses morphologically connected operators. This algorithm does not have effective clutter removal technique.

[1] uses a multiresolution (3 levels) motion parameter estimation technique, followed by motion compensated differencing of high frequency components obtained by Discrete Wavelet Transform (DWT) to isolate the objects. This process is computationally inexpensive, but suffers from quality problems in the presence of global motion and background clutter. These quality problems can be attributed to the fact that [1] does not deal with the inherent shift variance problem associated with discrete wavelet transform.

It is well known that while applying wavelet transform to a function, if the function and wavelets are discrete, the samples of the translated function, say, $f(x+n)$ do not correspond with the translated coefficients representing the discrete wavelet transform unless the translations of the form $n = k2^p$ (where k and p are integers) are considered. As [1] uses three levels of DWT decomposition, this problem may increase with each level. Thus, the compensation process done on the high frequency wavelet components is significantly affected, leading to improper segmentation.

In this paper, we contribute a new moving object segmentation algorithm for tracking application that successfully handles difficult situations created by background clutter and global motion, maintaining the low complexity level of the algorithm proposed in [1]. For this, we consider both the high and low frequency information obtained from a single level Stationary Wavelet Transform (SWT). We use stationary wavelet transform to avoid the inherent shift variance problem present in discrete wavelet transform.

As we know that using stationary wavelet transform increases the computational load when compared to discrete wavelet transform, we propose new processing steps which maintain the low complexity level and enhance the segmentation performance. The processing steps consists of motion adaptation, clutter removal and region creation techniques. We achieve this low complexity level by working on binary images during clutter removal and region creation.

Motion adaptation is done by using an iterative technique to estimate the motion parameters in accordance to the global motion present. Binary median filtering acts as an effective clutter removal tool. Region creation consists of a novel gap filling and false region reduction technique.

In section 2, we give an overview of the proposed segmentation algorithm. In section 3, 4, 5, 6, we explain the advantages of using both high and low frequency information,

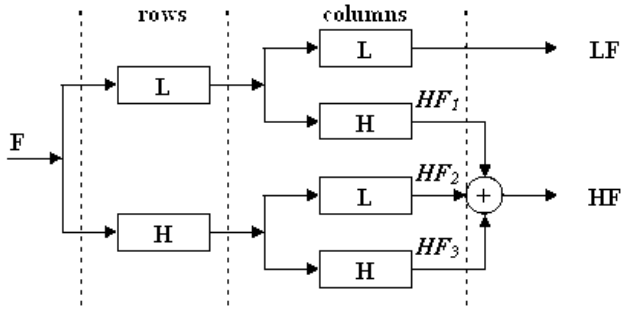
the motion adaptation, clutter removal and region creation techniques, respectively. In section 7, we present simulations to endorse the effectiveness of our algorithm. Here we compare the segmentation performance and execution time of the proposed algorithm to that proposed in [1]. Section 8 concludes this paper.

2. Overview of Proposed Algorithm

The proposed algorithm steps are depicted in Fig. 2. In the first step, single level SWT is applied on each frame, F, to obtain the high (HF) and low frequency (LF) information bands simultaneously as in Fig. 1. We use symlet wavelet filters (symN) to do the SWT as symlet family of filters are symmetric and hence have linear phase response [8]. HF_1, HF_2 and HF_3 consist of the horizontal, vertical and diagonal edge information of the frame F, respectively. In each frame, the three high frequency images, HF_1, HF_2, HF_3, are added to give the high frequency band, HF, as follows

$$\begin{aligned} \left[\text{ LF } HF_1\ HF_2\ HF_3 \right] &= SWT \left[\text{ F } \right] \\ \text{HF} &= HF_1 + HF_2 + HF_3 \end{aligned} \tag{1}$$

Thus, HF will have all the high frequency edge information present in the frame F.



where L and H are lowpass and highpass Wavelet Filters

Figure 1. Two-Dimensional SWT Decomposition

In the second step (Sec. 4), initial motion parameters representing camera motion are estimated by solving an error minimization problem obtained by 2D optical flow criterion [7], applied to the low frequency band. Eight parameter projective/bilinear model [6] is used to represent these motion parameters. To solve the 2D optical flow equation, the partial derivatives (i.e., the horizontal, vertical and temporal edges) E_x, E_y and E_t are estimated as in Eq. 2. Differential of Gaussian (DoG) is used to detect the horizontal and vertical edges of the LF.

$$\begin{aligned} \left[\ E_x\ E_y \ \right] &= DoG \left[\ LF_{F_c} \ \right] \\ E_t &= LF_{F_c} - LF_{F_p} \end{aligned} \tag{2}$$

where LF_{F_c}, LF_{F_p} are the low frequency bands of the current, F_c, and previous frame, F_p, respectively.

In the third step, motion adaptation is achieved by calculating the final motion parameters using an iterative approach (Sec. 4). The estimated final motion parameters are then used to predict the current frame from the previous frame in each iteration. Then, the difference between the predicted and the actual frame is calculated for both the high, HF, and low frequency, LF, bands. These two difference images are then thresholded to give the binary high, B_{HF}, and low, B_{LF}, frequency images, respectively. We use a simple frame difference technique to avoid large computations involved in other complex background estimation and subtraction techniques. The resulting two binary images are added to obtain the current binary image B_c.

$$B_c = B_{LF} + B_{HF} \tag{3}$$

The thresholds to obtain B_{LF} and B_{HF} are computed using the histogram distribution of the gray levels in the difference images.

In the last step, clutter is removed and regions are created. Median filter (Sec. 5) is applied to B_c to remove clutter. This is followed by a gap filling algorithm to fill gaps (black (0) pixels) within concentrated clusters of white (1) pixels in B_c (Sec. 6.1). A false region reduction technique (Sec. 6.2) is then used to reduce false regions. As we work on binary images less amount of computation is required. The performance of the proposed segmentation algorithm is shown in Sec. 7. The segments are in the form of rectangular shape which are usually used in various rule-based tracking algorithms, for example, as proposed in [1].

3. Processing both HF and LF Bands

The wavelet transform is able to retain the spatial characteristics of a signal in the transform domain in the form of edges and can thus represent objects. In general, object information is significantly present in both HF and LF wavelet bands. We, therefore, consider both HF and LF band information for object isolation. LF band information is especially useful in blurred video sequences, due to incorrect focusing. Fig. 3(d) and Fig. 4(d) show the amount of information acquired by using both LF and HF bands. Note in Fig. 3(b) and Fig. 4(b) the object information is not clearly present when only the HF band is considered.

4. Motion Adaptation

Simulations have shown (Sec. 7) that the use of three levels of DWT [1] is inaccurate in case of large global motion. Furthermore, three levels of DWT unnecessarily increases computational complexity when global motion is small. It is also noted that the shift variance problem of DWT further worsen the segmentation performance. To overcome this

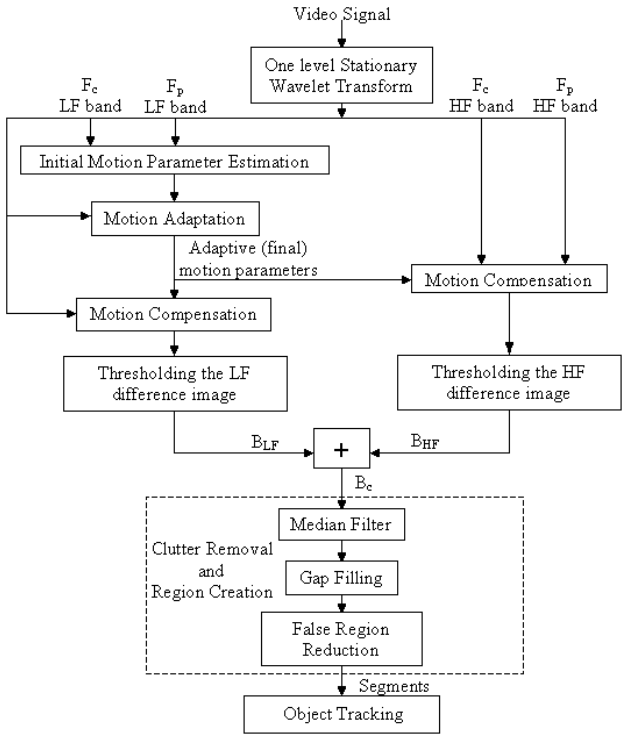


Figure 2. Block diagram of the proposed algorithm.

drawback, we propose to use a single level SWT followed by an iterative estimation of motion parameters. We use an eight parameter bilinear/projective two-dimensional motion model (Eq. 4, [6]).

$$q = \left[\; q_0 \; q_1 \; q_2 \; q_3 \; q_4 \; q_5 \; q_6 \; q_7 \; \right]^T \tag{4}$$

Using LF band, the initial motion parameters are first estimated. For this, we use the optical flow minimization technique [6], which is a computationally less expensive way to estimate motion vectors with certain loss of accuracy.

Motion adaptation is an important requirement while determining motion parameters. Since the initial motion parameters (Eq. 4) are fractional approximation of the actual values, we propose a motion-adaptive proportional increment of the initial value (q) which leads to parameters close to the actual motion parameters. The increment is done in an iterative way and the number of iterations used is dependent on the amount of the estimated global motion.

This concept is implemented by comparing the mean square error between the predicted and the current frame in each iteration (MSE_c) with that of the previous iteration (MSE_p). The iteration will continue as long as MSE_p is greater than MSE_c to a precision of five decimal places. Note that in each iteration we are not estimating the motion parameters but are proportionally incrementing them. With this approach, there is no significant increase in computational load.

5. Clutter Removal

Clutters are the non-target regions that causes false alarm in the background, or obscure the target objects and thus deteriorates the performance of object segmentation. In addition, clutter can result from inaccurate processing steps such as motion compensation. In particular, clutter might be present in frames with high frequencies in the background. These clutters are represented as sparse white (1) pixels in the binary image.

We use median filter, which effectively removes these unwanted white (1) pixels in the binary image. We adapt the size (w_rxw_c) of the filter window to the size of the input frame. In general, the number of neighboring white (1) pixels representing background clutter will be more for a large frame. Thus, the window size of the median filter and size of the frame are linearly related as in Eq. 5.

$$w_r = round\left(\frac{f_r}{2^7}\right) + 1, \;\; w_c = round\left(\frac{f_c}{2^7}\right) + 1 \tag{5}$$

where f_r and f_c are the number of rows and columns in the frame, respectively. For video sequences in CIF and NTSC/PAL format, for example, we use a window size of 4x3 and 7x6, respectively. Depending on the size of the median filter, its output may be fractional, i.e., neither a black (0) nor a white (1) value. We convert thus all fractional values to white (1).

As can be seen in Fig. 3(d) and Fig. 4(d), large amount of clutter is present in the background. The proposed filter effectively removes the clutter (Fig. 3(e) and Fig. 4(e)) which in turn reduces the creation of false objects.

6. Region Creation

6.1. Gap Filling

A novel gap filling technique is used to connect certain white (1) pixels (by replacing black (0) with white (1) pixels) in the binary image. For this, the binary image is scanned first horizontally, and the number of black (0) pixels between two white (1) pixels are counted. If the counter is not greater than a threshold, then, a gap is assumed and all the black (0) pixels are replaced by white (1) pixels in the detected gap. Similarly, the binary image is then scanned vertically using a different threshold. The thresholds for the horizontal and vertical scan are set based on object and frame sizes. As can be seen in Fig. 3(f) and Fig. 4(f), the proposed gap filling satisfactorily creates regions at the correct object position.

6.2. False Region Reduction

To reduce false regions in the binary image resulting from the previous algorithm steps, we (1) extend the iterative false object reduction [1] to binary filled images (to significantly reduce the number of iterations), (2) propose adaptation of thresholds used (to reduce ambiguities), and (3) propose "lump" processing (to reliably detect false regions). For this, both the rows and columns of the binary image are iteratively processed as follows.

To process the rows, all the pixels in a particular row are first added to give the value sr. This is then repeated for each row of the binary image resulting in an array of sr, called *column vector* CV. All $sr \in CV$ below a certain threshold are set to zero. This threshold varies from frame to frame and is set to be a percentage of the maximum $sr \in CV$.

A group of consecutive non-zero $sr \in CV$ is called "lump". If the length of a "lump" is greater than a predefined threshold, then, it is assumed to represent at least one object. Otherwise all sr in the "lump" are set to zero. Then, only the rows corresponding to the non-zero $sr \in CV$ are considered for further processing. This results in smaller matrices within the binary image. Note that one lump corresponds to one matrix and a matrix represents at least an object. Hence if there is more than one lump in the CV satisfying the predefined threshold, multiple matrices are produced.

Similarly, the values in each column of the resulting binary matrices are processed to give an array called *row vector* RV. If the length of a "lump" in RV is not greater than predefined threshold, the elements representing this "lump" in RV are set to zero. Then, only the columns corresponding to the non-zero elements of RV are considered, which generates again smaller matrices.

The proposed column and row vector processing is applied *iteratively* on the smaller matrices obtained from previous iteration until the size of each of the matrices does not change. Each matrix then, will represent an object. With this, regions created at positions where there is no object, are effectively removed. Note that the gap filling step creates prominent regions at the object positions compared to background, hence, it reduces the number of iterations required for false region reduction.

7. Experimental Results

We have tested the proposed algorithm and compared it with the method in [1]. We present here sample results for the following video sequences (where F(n) denotes the n^{th} frame or field):

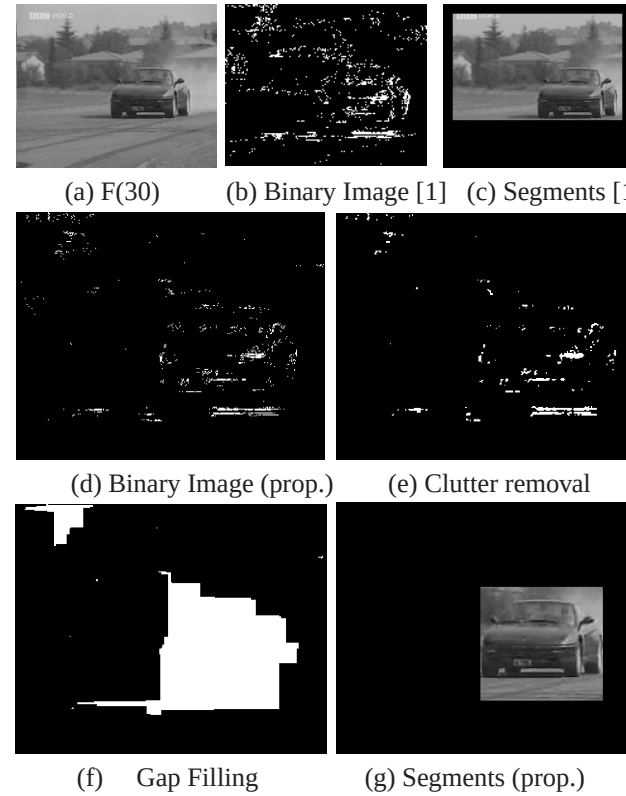
1. One Car Sequence (OCS) of size 360x180: OCS includes background clutter. It has a relatively small object entering and leaving the scene in the presence of less global motion.

2. Ferrari Sequence (FRS) of size 352x288: FRS consists of high global motion, e.g., zoom out and translation. There is smoke in the background acting as camouflage.

3. Coast Guard Sequence (CGS) of size 352x240: CGS includes two objects (where one disappears), significant global motion, and background clutter (e.g., water and rocks).

4. Multiple Object Sequence (MOS) of size 320x240: MOS includes moving objects like vehicles and people, entering and leaving the scene. This sequence has very less global motion with objects taken at an intersection of roads.

Figure 3 shows intermediate processing stages for FRS. The proposed algorithm (prop.) successfully combines HF and LF bands (Fig. 3 (d)), reduces the background clutter such as camouflage (Fig. 3 (e)), and creates region (Fig. 3 (f)). As can be seen, it provides more accurate segments compared to [1].

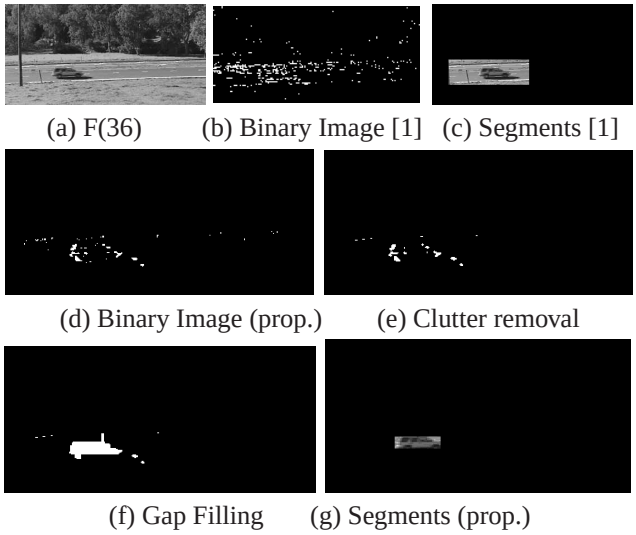


(a) F(30) (b) Binary Image [1] (c) Segments [1]

(d) Binary Image (prop.) (e) Clutter removal

(f) Gap Filling (g) Segments (prop.)

Figure 3. Intermediate segmentation stages in F(30) for FRS.



(a) F(36) (b) Binary Image [1] (c) Segments [1]
(d) Binary Image (prop.) (e) Clutter removal
(f) Gap Filling (g) Segments (prop.)

Figure 4. Intermediate segmentation stages in F(36) for OCS.

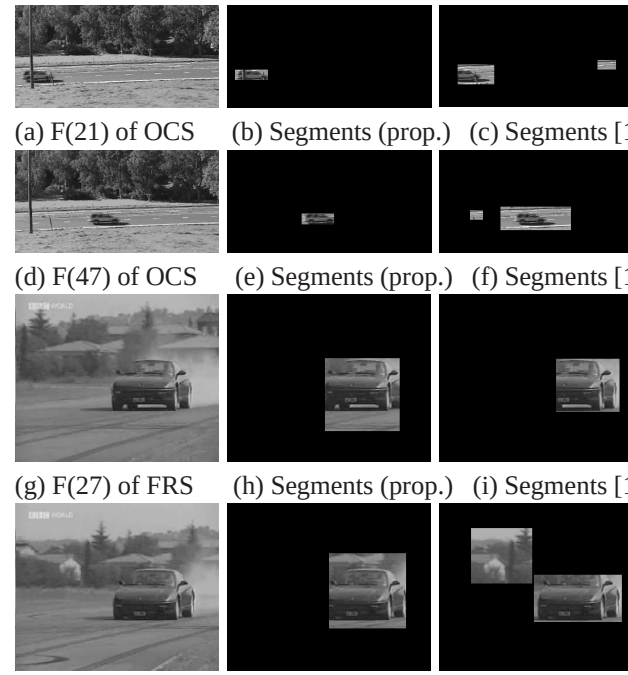
Figure 4 shows the various intermediate segmentation stages for OCS. As can be seen, the proposed algorithm results in significantly more effective segmentation as compared to [1], even in the presence of heavy clutter.

It can be noted that due to the shift variance problem in DWT, we observe false motion in Fig. 3(b), Fig. 4(b) at areas where no moving object exists. Similarly, use of DWT may produce no motion at the moving object positions. SWT significantly tackles this problem (Fig. 3(d), Fig. 4(d)) as it is shift invariant.

Figure 5 summarizes the good performance of the proposed algorithm compared to [1], as the later produces false segments due to its inability to deal with high global motion and background clutter.

Figure 6 shows the improvement in performance of the proposed algorithm as compared to [1] in CGS. In this sequence, trails on water by the boat, rocks and trees are present significantly in HF bands. [1] uses DWT (which is shift variant) and HF information only for segmentation of objects and hence results in false object and improper segmentation in various frames. Use of SWT along with iterative approach to estimate motion parameters and considering both HF and LF information, results in effective segmentation in our proposed algorithm.

Table 1 compares the complexity level of our algorithm to that in [1], in terms of average execution time (seconds per frame). Note that in OCS, there is little global motion. Hence, the number of iterations required per frame in the motion adaptation are small and the complexity of our algorithm is less than in [1]. Note that in FRS, there is large global motion, clutter and camouflage. Due to large global



(a) F(21) of OCS (b) Segments (prop.) (c) Segments [1]
(d) F(47) of OCS (e) Segments (prop.) (f) Segments [1]
(g) F(27) of FRS (h) Segments (prop.) (i) Segments [1]
(j) F(35) of FRS (k) Segments (prop.) (l) Segments [1]

Figure 5. Segmentation in OCS and MOS.

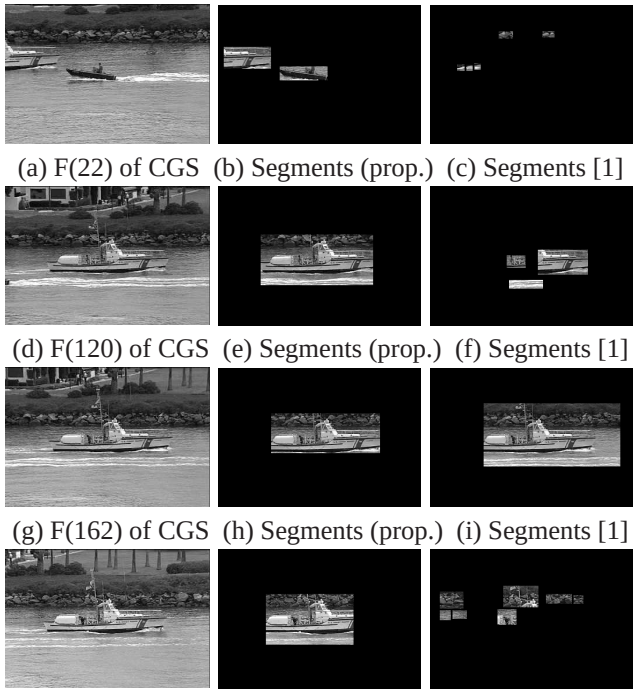
motion, the number of iterations required per frame in the motion adaptation increases. However, our algorithm maintains the same complexity level as in [1] with better segmentation. In CGS, our algorithm outperforms the algorithm proposed in [1], in segmentation of moving objects, whereas the execution time per frame is approximately the same. In MOS, our algorithm is faster than that proposed in [1], where moving objects in traffic are tracked.

Test sequence	sec/frame (prop.)	sec/frame ([1])
OCS	1.66	2.12
FRS	2.71	2.48
CGS	2.41	2.38
MOS	1.98	2.24

Table 1. Computational complexity comparison between the proposed (prop.) and the reference [1] algorithm (implemented using the C language under Sun OS 5.8 with 1 GHz processor).

Figure 7 demonstrates successful tracking of objects using our algorithm where gray levels are used to label the objects during tracking. MOS is an example for traffic monitoring application.

Note that the inconsistent segmentation by the algorithm proposed in [1] might significantly complicate object tracking.



(a) F(22) of CGS (b) Segments (prop.) (c) Segments [1]
(d) F(120) of CGS (e) Segments (prop.) (f) Segments [1]
(g) F(162) of CGS (h) Segments (prop.) (i) Segments [1]
(j) F(226) of CGS (k) Segments (prop.) (l) Segments [1]

Figure 6. Segmentation comparison with CGS.

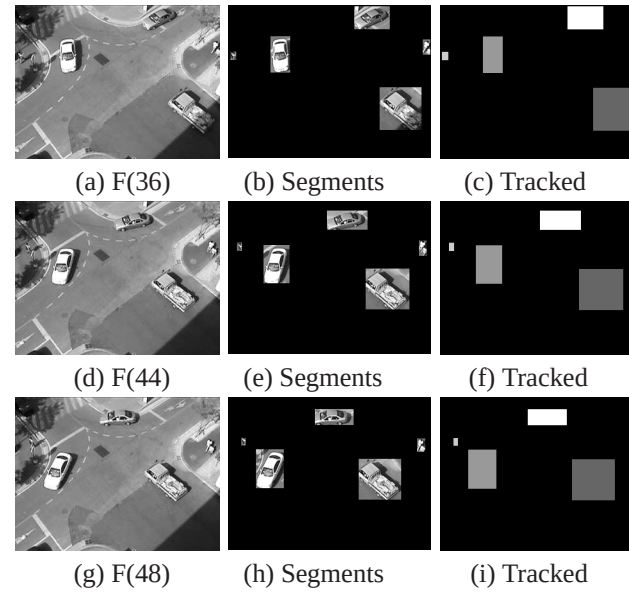
8. Conclusions

In this paper, we have proposed a new moving object segmentation algorithm that operates on high and low frequency bands obtained by stationary wavelet transform simultaneously. It is effective in the presence of (1) small or large global motion, (2) cluttered background, and (3) camouflage. The proposed segmentation uses novel motion adaptation, clutter removal and region creation techniques. The motion adaptation technique is able to adapt to variation in global motion. The proposed clutter removal and region creation technique effectively reduces the possibility of false object segmentation. Comparison with related work shows that the proposed algorithm is significantly more reliable, particularly in reducing false object regions with low computational complexity.

In our future work we shall investigate on making thresholding spatiotemporally adaptive. We are also contemplating to incorporate a fast shadow detection algorithm and removal to furthermore improve the results.

References

[1] Yiwei Wang, J.F. Doherty,R.E. Van Dyck, "Moving object tracking in video," in *Applied Imagery Pattern Recognition Workshop, 2000. Proceedings. 29th*, Oct. 2000, pp. 95–101.



(a) F(36) (b) Segments (c) Tracked
(d) F(44) (e) Segments (f) Tracked
(g) F(48) (h) Segments (i) Tracked

Figure 7. Segmentation and tracking in MOS.

[2] X. Cohen, G. Medioni, "Detecting and tracking moving objects for video surveillance," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, June 1999, vol. 2, p. 325.

[3] Ulisses Braga Neto, J. Goutsias "Automatic Target Detection and Tracking in Forward-Looking Infrared Image Sequences using Morphological Connected Operators," *33rd Annual Conference on Information Sciences and Systems - CISS'99*, March 1999, Vol. I, pp. 173-178.

[4] D. Wang, "Unsupervised Video segmentation Based On Watersheds And Temporal Tracking," in *Circuits and Systems for Video Technology, IEEE Transactions on*, Sept. 1998, vol. 8, pp. 539–546.

[5] Y.John, A. Wang, Edward H. Adelson, "Spatio-Temporal Segmentation of Video Data," Tech. Rep. No. 262, The MIT Media Laboratory, MIT, 1994.

[6] Yao Wang, Joern Ostermann, Ya-Qin Zhang, *Video Processing and Communications*, Prentice Hall, 2002.

[7] S. Lertrattanapanich, N. K. Bose, "Latest Results On High-resolution Reconstruction From Video Sequence," *Technical Report of Institute of Electronics, Information and Communication Engineers*, pp. 59–65, Dec. 1999.

[8] Ingrid Daubechies, *Ten Lectures On Wavelets*, Society for Industrial and Applied Mathematics , 1992.