Robust Algorithm for Object Detection and Tracking in a Dynamic Scene

Saad A. Yaseen and Sreela Sasi
Department of Computer and Information Science, Gannon University, Erie, Pennsylvania, United States of America
Email: s.iq88@yahoo.com, sasi001@gannon.edu

***Abstract*—The main research challenge for a security and surveillance system is to create a real-time fully autonomous system that is also robust. In this research, a robust approach for real-time object detection and tracking in a dynamic scene using a moving camera is presented. The detection of the moving object and the tracking of the detected object are accomplished using a modified version of the enhanced SURF algorithm. This includes a color feature also to achieve a more accurate and robust results. This approach is able to track the detected object while reentering the scene after being absent for a short period of 4 or 5 frames. The regular SURF, enhanced SURF, and the current approach are implemented and the results are compared for speed and accuracy.**

***Index Terms*—object detection, speeded-up robust features (SURF), scale- and rotation- invariant, object tracking**

I. Introduction

In today's world security and surveillance are extremely important in dynamic environments due to global security issues. Automatically understanding events in a scene has been the video surveillance system's goal. Moving object detection and tracking in consecutive images has been a very important problem in the field of computer vision research including vehicle detection and identification, abnormal behavior detection, crowd control, access control, and crime prevention. Currently, the surveillance systems are offline and require a lot of memory space to archive the video streams that eventually will be monitored by a human operator [1]. The surveillance systems can be manual, semi-automatic, or fully autonomous. The manual system requires human observation for many uneventful hours. Semi-automatic video surveillance systems include some computer processing to highlight some of the important events in the scene and analysis of the events by humans. The fully autonomous system requires low-level processing from input video and does high-level decision-making tasks.

There are many types of video surveillance systems. The surveillance systems can be manual, semi-automatic, or fully autonomou[2]. The manual system requires human observation for many uneventful hours. Semi-automatic video surveillance systems need computer processing to highlight the important events in the scene and then a human to analyze the events. The fully autonomous system performs low-level processing from input video and does high-level decision-making tasks such as abnormal behavior detection, traffic control, gesture identification, object detection [3], etc. Online processing algorithms interpret the video sequence in real time and hence reduce the storage requirements for archiving the videos. Feature extraction is used to simplify the amount of resources required to describe a large set of data accurately and makes it easier for real-time applications. Different techniques of feature extraction were presented in [4].

Speeded-Up Robust Features (SURF) algorithm is one of the best approaches for feature extraction and is suitable for real-time applications [5]. The processes of SURF consist of three steps: detection of the interest points regardless of the viewpoint; description that is unique to each feature point and does not depend on the feature scales and rotation; and matching the predetermined interest points between consecutive frames. The SURF algorithm is used to detect and track a moving object in a dynamic scene without any human interaction. It detects many interest points when the image frame is very complex. These detected features result in a heavy computation burden for the subsequent processes.

The enhanced SURF algorithm reduces the computational complexity of SURF, and exhibits an efficient performance [6]. The enhanced SURF algorithm reduces the number of the detected feature points by changing the range of the non-maximum suppression. Hence, only the strong features are selected. Also, enhanced SURF minimizes the repeated calculation needed for calculating the feature point's dominant orientations. The RANdom SAmple Consensus (RANSAC) technique is used to remove the outliers when matching the feature points between adjacent frames. The RANSAC is a re-sampling technique that generates candidate solutions by using the minimum number of observations required to estimate the model parameters [7]. The non-maximum suppression is a post-processing method for eliminating redundant object detection windows by setting a threshold value [8].

The regular SURF and the enhanced SURF work on gray scale images. In this research a modified version of the enhanced SURF is proposed that has included a color feature also to achieve a more accurate and robust result. With this algorithm the object will be tracked accurately even after it disappears from the window for a short

Manuscript received January 20, 2014; revised April 28, 2014.

doi: 10.12720/joig.2.1.41-45

period of 4 or 5 frames. Also in this research, the algorithms for regular SURF, enhanced SURF, and the proposed approach are compared for their speed and time.

The rest of this paper is organized as follows: Section II presents the background research related to object detection and tracking. Section III explains the proposed approach. Section IV provides the simulation results. The conclusion and future work is presented in section V which is followed by the references.

II. Background Research

Moving object detection is the first low level important task for any video surveillance application. Detection of a moving object in a dynamic scene is a very challenging task. Tracking is required in higher level applications that require the location and the shape of the object in every frame. In this section different approaches have been explained for detection and tracking.

A. *Object Detection Methods*

The main approach for object detection using a stationary camera is by maintaining the background as an average of the frame sequences. The detection of the foreground objects is achieved by determining differences between the object in the current frame, i and the model of the background, if or this frame as shown in equation (1) [9].

$$|\mathrm{Frame}_\mathrm{i} - \mathrm{Background}_\mathrm{i}| > Threshold \qquad (1)$$

The pixels are classified as foreground if their values are greater than a specified threshold value. It is a difficult process to specify a proper threshold value. If the threshold value is high, some of the foreground pixels will be missing and classified as background. Background subtraction is a good approach to detect moving objects from video frames captured using a stationary camera. Background subtraction fails if the illumination drastically changes in consecutive video frames. Also, this method will fail when there is a temporarily moving object such as leaves of a tree when recorded outdoors.

Temporal differencing is another way to detect moving objects, although it requires temporarily saving the sequence of consecutive frames to reduce the number of false negatives [3]. This saved information indicates that the region has changed dramatically in the consecutive video frames. The temporal differencing provides a good result for dynamic scene changes; however the approach fails to detect objects in the scene if the objects stops moving. This happens because the temporal differencing fails to detect any differences in the pixels between consecutive frames.

Paragios and Derichepresented a framework for detecting and tracking multiple moving objects in image sequences using a mixture model that consists of two components: the static (background) and the mobile(moving objects). Both components are of zero-mean and obey Laplacian or Gaussian law. This statistical framework is used to provide the motion detection boundaries. The first frame is used to provide the object detection boundaries. Then, the detection and the tracking problems are addressed in a common framework that employs a geodesic active contour objective function. This function is minimized using a gradient descent method, where a flow deforms the initial curve towards the minimum of the objective function, under the influence of internal and external image dependent forces. Using the level set formulation scheme, complex curves are detected and tracked while topological changes for the evolving curves are naturally managed. The Hermes approach was used to reduce the cost of the implementation [10]. The Hermes algorithm combines the Narrow Band and the Fast Marching method by employing the idea of a selective propagation (Fast Marching) over a relatively small window (Narrow Band).They used a smart Narrow Band method that uses ideas from Fast Marching (e.g., fastest pixel) method and resulted in a drastic decrease of the required computational cost. Thus, at each step, this approach selects the pixel of the front, preserving the highest absolute propagation velocity, and performs a local evolution to the level set frame within a circular window centered on this pixel. This method is used in the current approach for object detection.

B. *Target Tracking Methods*

Point tracking is one of the approaches to track objects. Point tracking represents an object as points in each frame and tracks these points in the consecutive frames. Point representation of objects can be classified into deterministic and statistical methods. The deterministic method depends on velocity and common motion to match point correspondence [2]. On the other hand, the statistical method uses position and size of the object. Kalman filter [11] and a non-Gaussian state-space approach to the modeling of non-stationary time series [12] are representative models of the statistical point tracking.

The kernel tracker identifies the target object with a primitive object shape such as rectangle. Tracking is computed by calculating the object motion between consecutive frames. There are two types of kernel tracking: template model and appearance model. The template model is known for computational simplicity. Collins *et al.* performed target tracking using sub-sampling method with motion estimation [13]. This minimizes the template matching process computations. On the other hand, appearance model tracks an object by computing the eigenvector of the affine transformation.

The Contour Tracking method identifies the target object in the next frame using the outline contour from the previous frame. This approach is known for its ability to track objects with complex shapes and recognize changes to the shape over time. Peterfreund proposed a new active contour model for people tracking based on Kalman filter and spatio-velocity space [14].

III. The Proposed Approach

The proposed method adds a color feature to the enhanced SURF algorithm to make it more robust. Video clippings of road traffic recorded using a camera mounted on moving vehicle are used for this research. First, the frames of the input video are read. The color feature of the object is identified and stored by converting the color from RGB to YCbCr color space. Then the feature points of the moving object are detected and extracted between the adjacent frames using the enhanced SURF detector. The interest points are then surrounded with a rectangle box. This box is sent with its coordinates to the enhanced SURF matching algorithm to find a match in the subsequent frames.

The enhancement of SURF reduces the number of the detected interest points. Also, it limits the approach to detect stronger features by changing the window size for the range of the non-maximum suppression. The non-maximum suppression was applied in 7 x 7 x 3 neighborhood to reduce the number of feature points detected and to simplify the calculations in this approach. The interest points are the points that are the extrema among 48 neighbors in the current level and the 2 x 49 in the level above and below in an octave.

Sliding window is the method that is used to calculate the orientation of a feature point. The sliding window covers an angle of 60 ° by shifting around a circular region. The Haar wavelet is calculated inside the circular window for the horizontal dx and vertical dy responses. The two summed vectors determine the orientation of the feature point. The sliding window shifting step is 5° and that produces many overlapped regions in which sum of the response are calculated repeatedly. For instance, assume that the first sliding covers $0-60\,°$ and calculates the Haar wavelet for the region. Then the next shift is for 5-65 degrees that yield an overlap of $5-60\,°$degree regions for which the response is already calculated. Calculation of the sum of horizontal and vertical Haar wavelet responses respectively at each degree for (0-360) and store them in x[360] and y[360]. Calculate the integral of X[i] and Y[i] respectively, denoted by Dx[i] andDy[i] as given in equation (2).

$$D_x[i] = \begin{cases} X[0] & i = 0 \\ D_x[i-1] + X[i] & i \in (0, 360) \\ D_x[i-360] & i \epsilon\, [360, 420] \end{cases} \qquad (2)$$

The same process works for calculating, $D_y[i]$.

The second step is to calculate the Haar wavelet responses in the 60 degree sensor region using equation (3).

$$\mathrm{sum}_x[i] = D_x[i] - D_y[i-60]$$

$$(i \in [60, 420]) \qquad (3)$$

After calculation of the $\mathrm{sum}_y[i]$ by following the same step for $\mathrm{sum}_x[i]$, the local orientation of the vector[i] is

$$\mathrm{vector}[i] = \begin{pmatrix} \mathrm{sum}_x[i] \\ \mathrm{sum}_y[i] \end{pmatrix}$$

The length of the local vector is given by equation (4).

$$|\,\mathrm{vector}[i]| = \sqrt{\mathrm{sum}_x[i]^2 + \mathrm{sum}_y[i]^2} \qquad (4)$$

Then choose the one vector with the maximum length. The dominant orientation of the feature points is the longest local orientation vector over all the windows. The shifting step for the sliding window is calculated at each degree.

The global motion estimation [15] method is used in regular SURF to match the feature points from the consecutive frames. The enhanced SURF [6] detects feature points in the rectangle area only, not the entire frame. These features are then matched with the key points extracted from subsequent frames. This makes the matching method quicker than the regular SURF [5]. The RANdomSAmple Consensus (RANSAC) is used to validate the matching process and remove the invalid matching points to get the best of the interior points [7]. If a match is found, then the algorithm updates the coordinates of the rectangle area, and is passed to enhanced SURF algorithm to track the right object. This algorithm can track the object accurately even after it disappears from the window for a short period of 4 or five frames. This is possible because the algorithm will be checking for the color component of the object in the next frame. The architecture of the proposed approach is shown in the Fig. 1.

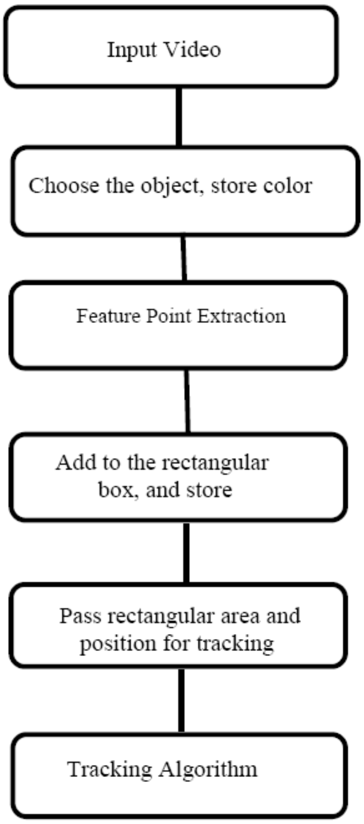

Input Video
Choose the object, store color
Feature Point Extraction
Add to the rectangular box, and store
Pass rectangular area and position for tracking
Tracking Algorithm

Figure 1. Architecture for the proposed approach

The flowchart of the tracking algorithm is given in Fig. 2.

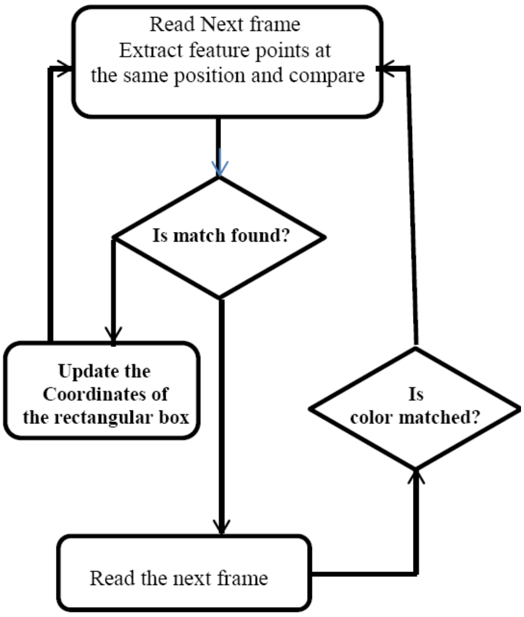


Figure 2. Flowchart for the tracking algorithm

IV. Simulation Results

The proposed modified version of enhanced SURF algorithm was tested with five video clippings from YouTube for traffic videos of durations of 10, 15, 16, 20, and 25 seconds. The results showed that the algorithm detected the object (car) and provided the information for tracked vehicle. Users may select one of the cars by surrounding it with a rectangular box. Then the feature points are found using the enhanced SURF algorithm and as shown in Fig. 3. From these feature points, only the strongest feature points of the rectangular bounding box are found using the RANSAC non-maximum suppression technique.

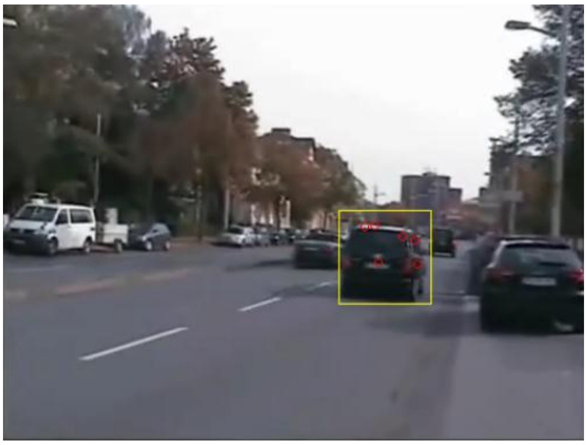


Figure 3. Detected feature pointsthat are kept inside the bounding box

The correspondence of the strongest feature points inside the rectangular bounding box are searched only in the corresponding region of the next frame. This works because there will not be much advancement of moving pixels inside the bounding box between adjacent frames. The invalid points are removed using the RANdom SAmple Consensus (RANSAC) algorithm. This is shown in Fig. 4.

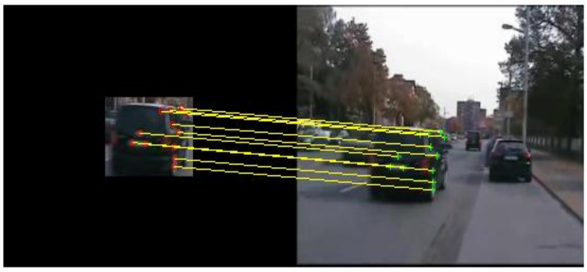


Figure 4. Matching features points from adjacent frames

The affine transformation is used to update the location of the moving object and the bounding box coordinates. The proposed algorithm was able to track the object until the last frame by matching the feature points of the vehicle in the consecutive frames. The simulation was done using regular SURF, enhanced SURF, and the proposed modified version of enhanced SURF and the results are tabulated in Table I. The computer that is used to test the algorithms has 2.13 GHz processor and 2 GB RAM. The time computed is only for detecting the features and matching them in subsequent frames. The result indicates that enhanced SURF takes much less time compared to regular SURF. But, it takes slightly less time than the current approach.

TABLE I. The Simulation Results

Algorithm	Time in seconds
Regular SURF	0.619607
Enhanced SURF	0.103402
The modified version of Enhanced SURF	0.11250

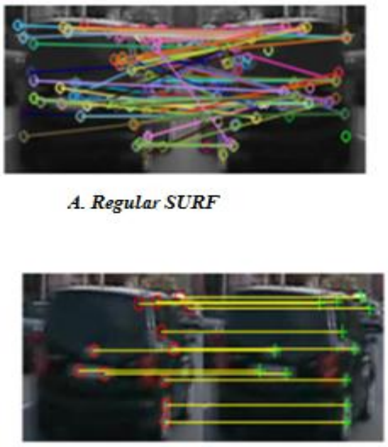
TABLE II. Results of Three Approaches for Different Scenarios

Scenario	Regular SURF	Enhanced SURF	Modified version of the Enhanced SURF
A car, C1 moving in a scene	detected	detected	detected
Two cars C1 and C2, moving in the same scene	Fail	Tracked the first car C1	Tracked the first car C1
The car C1 leaves the scene (out of the frame)	Fail	Fail	Started looking for the car C1
The car, C1 reenter the scene	Fail	Fail	detected
The car C3, with same characteristics of C1 with different color enter the scene	Fail	Fail	Not detected because C3 has a different color

The proposed algorithm was tested for five different scenarios: A car C1 moving in a scene, two cars C1 and C2 moving in the same scene, the car C1 leaves the scene (out of the frame), the car C1 re-enter the scene, and the car C3 with same characteristics (make and model) of C1 enters the scene. These five scenarios were tested using the regular SURF, enhanced SURF, and the current approach, and the results are tabulated in Table II.

Fig. 5 shows the results of using regular SURF, enhanced SURF, and the current approach. A large number of feature points are detected using regular SURF as shown in Fig. 5 (a). This will cause a heavy computation burden for the subsequent processes. Since the affine transformation requires only three pairs of matching features to achieve the image transformation, the weak feature points are eliminated for the enhanced and the current approach as shown in Fig. 5 (b). The number of the detected feature points is minimized by changing the rate of the non-maximum suppression, and hence only the stronger features are captured. Using this technique, the efficiency of the algorithm is improved.

The current approach is more robust since it can detect the moving object even after a short absence of 4 or 5 frames. This is possible because the color feature of the car is searched in the moving direction of the frame before applying the detection algorithm. This is more useful for a real-time application.



A. Regular SURF

B. Enhanced SURF and the current approach

Figure 5. Matching results using regular SURF, enhanced SURF and the current approach

V. Conclusion and Future Work

In this research, a robust method for moving object detection in a dynamic scene using a modified version of the enhanced SURF is presented. The videos are recorded using a moving camera. This method is able to track the selected object while re-entering the scene even after being absent for a short period of 4 0r 5 frames. This is possible because the color feature of the car is searched in the moving direction of the frame before applying the detection algorithm. This is more useful for a real-time application.

The regular SURF, enhanced SURF, and the current approach are implemented using the Computer Vision Toolbox of Matlab 2013a. The result indicates that enhanced SURF takes very less time compared to regular SURF. But, it takes slightly less time than the current approach. The time computed is only for detecting the features and matching them for subsequent frames. This approach can track a selected moving object from video even when multiple moving objects are present. The algorithm may be extended to track multiple moving objects in a video using a moving camera.

References

[1] L. Xiao and L. T. Qiang, "Research on moving object detection and tracking," in *Proc. 7th International Conference on Fuzzy Systems and Knowledge Discovery*, Hangzhou, China, vol. 5, 2010, pp. 2324-2327.

[2] I. S. Kim, H. S. Choi, K. M. Yi, J. Y. Choi, and G. Seong, "Intelligent visual surveillance–A survey," *International Journal of Control Automation and Systems*, vol. 8, no. 5, pp. 926-939, 2010.

[3] D. G. Thakore and K. A Joshi, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, 2012.

[4] A. Aichert. (January 9, 2008). Feature extraction techniques. [Online]. Available: http://home.in.tum.de/~aichert/featurepaper.pdf

[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, June 2008.

[6] J. Pan, W. Chen, and W. Peng, "A new moving objects detection method based on improved SURF algorithm," in *Proc. Control and Decision Conference (CCDC)*, Guiyang, China, May 25-27, 2013, pp. 901-906.

[7] K. G. Derpanis. Overview of the RANSAC algorithm. [Online]. Available: http://www.cse.yorku.ca/~kosta/CompVis_Notes/ransac.pdf

[8] A. Neubeck and L. V. Gool, "Efficient non-maximum suppression," in *Proc. 18th International Conference on Pattern Recognition*, vol. 3, 2006, pp. 850-855.

[9] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE International Conference* on *Systems, Man and Cybernetics*, vol. 4, Oct. 10-13, 2004, pp. 3099-3104.

[10] N. Paragios and R. Deriche, "Geodesic active contours and level sets for the detection and tracking of moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 266-280, 2000.

[11] M. H. Bahari, A. Karsaz, and H. Khaloozadeh, "A new algorithm based on combined fuzzy logic and Kalman filter for target maneuver detection," in *Proc. 1st International Symposium on Systems and Control in Aerospace and Astronautics*, Jan. 19-21, 2006, pp. 520-524.

[12] G. Kitagawa, "Non-gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1032-1041, 1987.

[13] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A system for video surveillance and monitoring," The Robotics Institute, Carnegie Mellon University, Pittsburgh PA, 1 The Sarnoff Corporation, Princeton, NJ 2000, CMU-RI-TR-00-12.

[14] N. Peterfreund, "Robust tracking of position and velocity with Kalman snakes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 6, pp. 564-569, 1999.

[15] H. Jia, M. Xie, and L. Ren, "An improved global motion estimation for practical objection detection," in *Proc. International Conference on Information and Automation*, 2008, pp. 1159-1162.

Saad A. Yaseen completed his MS in Computer and Information Science from Gannon University in December 2013. Saad is from Iraq.

Sreela Sasi is currently working as Professor in the Department of Computer and Information Science at Gannon University, Erie, Pennsylvania. USA. Sreela received Ph.D. in Computer Engineering from Wayne State University, Detroit, Michigan, USA in 1997. Research interests include Computer Vision, Data Analytics, algorithms for Decimal arithmetic, Reversible Logic and VLSI Design. She is a Senior Member of IEEE and a member of Eta Kappa Nu (HKN).