# AUTOMATIC VIDEO OBJECT SEGMENTATION AND TRACKING FROM NON-STATIONARY CAMERAS

Xuesong Le; Ruben Gonzalez
Institute for Integrated and Intelligent Systems, Griffth University,
Kessels Road, Nathan, QLD 4111, Australia
NICTA, 300 Adelaide St Brisbane, Qld, Australia 4000
X.Le@Griffith.edu.au

***Abstract*-This paper presents a robust algorithm for segmentation of moving objects in video, that first solves the global camera motion estimation problem and then processing the local object motion using the global motion parameters. This is work in progress.**

***Key Words*-Camera motion, global motion, extraction, segmentation**

## I. INTRODUCTION

The characterization of the visual content using the concept of video objects (VOs) is increasingly important in a variety of new multimedia applications such as content-based image retrieval, video editing, object-based compression and transmission and intelligent video surveillance. The introduced VOs roughly correspond to meaningful (semantic) content entities, such as persons, animals, buildings, or ships. VOs consist of regions of arbitrary shape with varying colour, texture, and motion properties. Such object-based representations provide new capabilities for accessing and manipulating visual information. For example, improved compression may be achieved by allowing the encoder to place more emphasis on objects of interest. Sophisticated content-based video analysis and retrieval can also be more effectively performed on video databases. The MPEG-4 standard which introduced the concept of VOs by specifying a general coding methodology for them, but it did not address the problem of VO extraction which remains a very interesting and challenging task.

Image and video object segmentation has always been a challenging task Although much work has been done in decomposing images into regions with uniform features, and especially for video data captured from fixed cameras, accurate and robust techniques for segmenting semantic video objects in general video sources captured from non-stationary cameras are used are still lacking. When the video source (or camera) is non-stationary the camera's self-motion needs to be first compensated for before further video analysis can be performed. Also, without relying on object recognition methods to identify image regions with specific semantic import, one is left with methods that identify regions that exhibit behaviors known to have semantic implications. Our goal is to develop a fully automatic, unsupervised method for real time video processing that can be applied to video surveillance, traffic monitoring, data compression for video conferencing and video editing.

In current state of art, segmentation of moving objects is classified into two groups; those based on optical flow versus change detection. The optical flow of a pixel is a motion vector represented by the motion between a pixel in one frame and its correspondence pixel in the following frame. Optical flow methods suffer from high computational complexity. Potter [1], Thompson [2], and Schunck [3] assume that all parts of an object have the same velocity. Optical flow estimation is often based on the movement of edges. The motion estimate for a given reference point on an edge is determined by the distance it travels relative to a superimposed grid. On the other hand, while change detection based method can achieve low computation complexity they are very sensitive to noise and small camera motion. Segmentation using this method begins with detecting changes between two successive frames to distinguish between temporarily changed and unchanged regions. Regions where motion has occurred are extracted as moving objects. Jain [4] used accumulative difference picture (ADP) technique to detect pixels belonging to actual moving regions. Ong [5], and Zhong [6], proposed various adaptive thresholding techniques in thresholding the difference. Lo [7] Wren [8], Stauffer and Grimson [9] proposed algorithms that construct the background images. These reported segmentation algorithms can reduce the high computation and problem and achieve high accuracy of extracting video objects captured from stationary cameras. This work reported in this paper focuses on natural video sequences captured from non-stationary cameras undergoing any affine motion transformation. We assume that the camera motion is consistent enough to ensure continuity of its motion with no rapid, haphazard movements. We do not consider changes induced by fast luminance changes. In comparison to the proposed approach, Farin [10], and Thakoor [11] used similar approaches to extract video objects with arbitrary rotational camera-motion. But the steps in estimating the camera motion parameters are different. Farin used least-squares solution to solve perspective camera motion parameters problem and a Least-Trimmed Squares regression algorithm in feature point selection. Thakoor used iterative weighted least square method to estimate affine motion

parameters. Structurally, this document is organized into three sections. The first section presents some background information necessary to place this work in context and facilitate the understanding of the research results presented herein. The second section presents research concepts and proposed algorithms. In the third section, an objective video object segmentation evaluation method is proposed.

## II. VIDEO OBJECT EXTRACTION

This section describes our approach for video object extraction and tracking. There are two main stages in proposed algorithm. In first stage, image registration is applied to every frame to compensate the camera motion.

In second stage, difference frames are generated and the moving objects are segmented out.

*Stage 1*

1. *Apply Harris Corner Detector to Frame $F_n$ and $F_{n-1}$.*
2. *Map selected corner set $C_n$ in $F_n$ with corner set $C_{n-1}$*
3. *Calculate the camera motion with 4 pairs of selected corners*
4. *Inversely transform $F_n$ with calculated camera motion parameter*
5. *Generate Difference frame $D_n$ between $F_{n-1}$ and transformed $\hat{F}_n$*

*Stage 2*

6. *Determine scene changes with difference frame $D_n$*
7. *If there is any scene change, go back to step 1 in stage 1, otherwise continues*
8. *Segment moving regions using the three frame method by evaluating union between region sets in difference frames $D_n$ and $D_{n-1}$.*
9. *Generate moving objects by removing shadows*

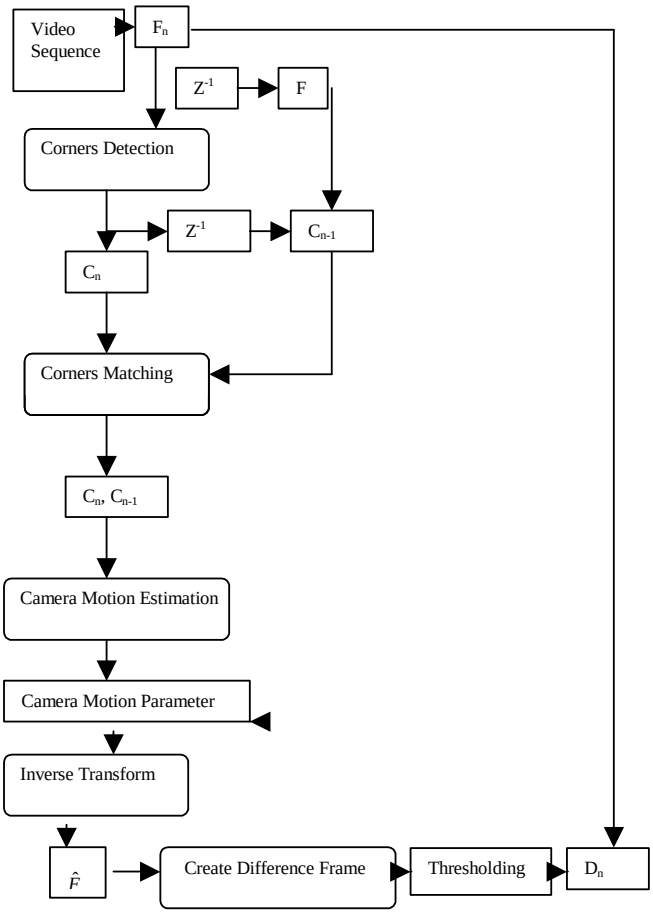


Figure 1. Image Registration in Stage 1.

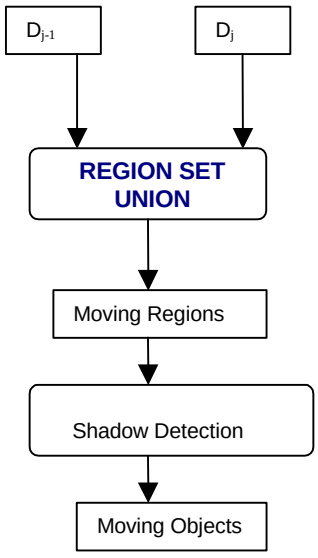


Figure 2. Moving Object Segmentation in Stage 2.

During the first stage, the Harris Corner detector is used as the feature extraction technique. The variation of the autocorrelation over different orientations is found by calculating functions related to the principle curvatures of the local autocorrelation. Harris Corner detector operates on the smoothed local structure matrix, C, which has the form

$$C_{Harris} = w_G(\sigma) * \begin{bmatrix} \left(\frac{\partial I}{\partial x}\right)^2 & \frac{\partial I \partial I}{\partial x \partial y} \\ \frac{\partial I \partial I}{\partial x \partial y} & \left(\frac{\partial I}{\partial y}\right)^2 \end{bmatrix} = w_G(\sigma) * \left( \begin{bmatrix} \frac{\partial I}{\partial x} \\ \frac{\partial I}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} \end{bmatrix} \right) \quad (1)$$

where $w_G(\sigma)$ is an isotropic Gaussian filter with standard deviation $\sigma$ and the operation * denotes convolution. A measure of the corner response at each pixel coordinates (x, y) is the then defined by

$$r(x,y) = \det(C_{Harris}(x,y)) - k(trace(C_{Harris}(x,y))^2, \quad (1)$$

where k is an adjustable constant and $C_{Harris}(x,y)$ is the 2x2 local structure matrix at coordinates (x, y).

Let $\lambda_1$, $\lambda_2$ be the two eigen-values of the matrix $C_{Harris}(x,y)$, $\det(C_{Harris}(x,y)) = \lambda_1\lambda_2$, and

$$\text{trace}(C_{Harris}(x,y)) = \lambda_1 + \lambda_2. \quad (2)$$

The most difficult step is matching of the corner candidates. The correspondence between the corners from the two frames must be established and the candidates having no counterparts should be rejected. Depending on the imagery this can result in thousands of corners needing to be matched which is time consuming. Furthermore these corners potentially belong to both static and moving objects in the scene. Ideally we limit the number of corners to be matched to around 100 to 200 of the more significant. Our approach uses the nearest neighbour method to find candidate corner matches between frames. The observed features of each detected corner $C_i$ include

- f1, the average distance or the mean between the corner and its neighbouring corners, which is the sum of all distances / the number of neighbouring corners.
- f2, the variance of those distance between the corner and its neighbouring corners
- f3, the two dimension skewness of the distance distribution relative to corner's two principal axes.
- f4, kurtosis is a measure of the "peakedness" of the distribution relative to corner's two principal axes.

By identifying those 4 neighbouring features, plus the x, y coordinates of the corner, and the orientation, we can have a complete set of corner feature descriptors {f1, f2, f3, f4, x, y, orientation). Thus for each corner $P_i$ found in $F_j$, we can infer how close the detected corner $C_i$ in $F_{j+1}$ is to be its closest match.

Then for each corner point $P_i$ in the frame $F_j$, an exhaustive search in its local neighbourhood is performed among all detected corners $\{C_i\}$ in frame $F_{j+1}$ to find its best match. The measure of closest match between corner $C_i$ and $C_j$ in two frames is based on the Euclidean Distance among all the features.

$$D(C_i, C_j) = \sum_{k=0}^{n-1} (f_k(C_i) - f_k(C_j))^2 \quad (3)$$

where n represents the number of features a corner $C_i$ can have, $f_k(Ci)$ is the value of feature in $K^{th}$ dimension. The best corner match is the one with minimum sum of Euclidean Distance among all the feature dimensions.

It is important that selected feature points do not belong to the moving objects but belong to the background. We solve this problem by the following method. First we assume that in any scene the screen area occupied by moving objects is less than the area occupied by the static background. Then for each feature point at $[x,y]^t$ within the block $B_i$ the previous frame, map the corresponding point at $[x`,y`]^t$ within the block $B_i$ in the current frame. The estimated motion vector for the feature point is $[(x-x`), (y-y`)]^t$. The magnitude $m_i$ of each motion vector is defined as:

$$m_i = \sqrt{(x-x')^2 + (y-y')^2}, \quad (4)$$

From all the pairs of feature points between corresponding blocks in two frames, select feature points which have the magnitude within one standard deviation of average magnitude. In such way, feature points selected come from the background and the feature points from the moving object are removed.

Once we have the registration points in two frames we use Wolberg's [12] approach to directly estimate the global camera motion. It first estimates the six affine transformation parameters for each split the block in the frames. Then the remaining two perspective parameters representing the distortion in x and y directions are derived by estimating the overall distortion of other six affine transformation caused by these two parameters.

In stage 2, the scene cut change can be easily detected if the total absolute pixel difference, t, in difference frame $D_j$, is greater than a certain threshold T. If the scene cut change is detected, camera motion parameters need to be calculated again. After scene changes have been processed, moving objects can be extracted by considering the set union from two difference images $D_j$ and $D_{j+1}$.

$$B_j = D_j \text{ AND } D_{j+1}, \quad (5)$$

The difference images represent a binary motion mask which distinguishes the changed and unchanged regions from frame $F_j$ to frame $F_{j-1}$. The changed regions found in difference frame $D_j$ are divided into moving regions and the uncovered background. Similarly, those changed regions found in difference frame $D_{j+1}$ are divided into moving regions and uncovered background as well. The set union

operation on two difference frames extracts the common moving regions and removes uncovered background.

For each pixel belonging to the shadow resulting from the segmentation step, the brightness component should change significantly in terms of absolute difference. The shadow mask SM [ 13 ] for each point p resulting from moving regions based on the following equations

$$SM(p)=\begin{cases}1 & if \alpha \le \dfrac{I(p).V}{B(p).V} \le \beta \qquad \alpha \in [0,1] \quad \beta \in [0,1] \\ 0 & otherwise \end{cases}, \quad (6)$$

where I(p) is the value of vector with R, G, B components, B(p) correspond to a point in the background, .V represent denotes brightness component in HSV space, which is transformed from RGB colour space. The lower bound α is used to define a maximum value for the darkening effect of shadows on the background and is approximately proportional to the light source intensity. The upper bound β prevents the system from identifying as shadows those points where the background was darkened two little with respect to the expected effect of the shadow.

## III. Evaluation Methodology

Video segmentation technology has received considerable attention to literature, and many segmentation algorithms have been proposed. However, very few comparative results of segmentation algorithms have been conducted. Currently, there are no standard, or commonly accepted, methodologies available for objective evaluation of image or video segmentation quality. Many researchers prefer to rely on qualitative human judgment for evaluation [14], [15]. This is a time-consuming and expensive process since there is no standard subjective video quality evaluation guideline for test environment setup and for how to score. Therefore, there is a need for an automatic, objective methodology both to allow the appropriate selection of segmentation algorithms as well as to adjust their parameters for optimal performance.

In this work, we will investigate quantitative performance measures for video object tracking and segmentation. The ground truth images are generated every 20 frames to compare the difference between segmented objects and reference objects. This method first compares the number of regions between the segmented images and ground truth images, where the regions in segmented images are generated with component labelling method. Then an edge pixel based method is applied to evaluate the difference between segmented objects and reference objects.

We use the property of pixels that lie on the boundary of regions to compare with the ground truth boundary. Specifically, we use the property of pixel location. The disparity evaluation of pixel location is based on chamfer matching [ 16 ]. By observation, overlaying a distance transformed image generated from ground truth image over its corresponding edge map generated from the segmented image, the extent of how well the segmented objects fit in terms of the pixel values in the distance transform map hit by the pixel in the segmented edge image is measured. In this way, the pixel distance between the segmented edge map and ground truth distance map can be computed. The total distance is the sum of the values of the pixels covered. Therefore, the likeliest match occurs when the sum of the pixel distance is the minimum. This minimum takes place when the segmented object boundary matches those in the reference distance map.

The precision-recall curve is a parametric curve that captures the accuracy and noise as the time varies. In disparity evaluation measure of pixel location, recall is the ratio of the number of relevant pixels found within the boundary of segmented object to the total number of relevant pixels in the boundary of target object. Precision is the ratio of the number of relevant pixels found within the boundary of segmented object to the total number of pixels within the boundary of segmented object. Both recall and precision are usually expressed as a percentage.

## IV. Conclusion

This paper investigates algorithms for automatic segmentation of moving objects in image sequences, which strives to achieve a set of goals. Firstly, this algorithm is to automatically and accurately segment and track moving objects from image sequence without any user interaction. Secondly it should be able to handle the camera motion, such as titling, rotation, panning and the fast moving objects as well. The shadow and ghost in each moving objects and image sequence can be automatically identified and removed. This algorithm can also adjust its camera motion parameter when there is a scene change. One of the aims is to provide a robust video segmentation algorithm which can be applied in real time application and adapt to different types of video sequences. Finally, this algorithm is evaluated in an objective and accurate way.

## References


[1 ] J.L.Potter, "Velocity as a Cue to Segmentation," in *IEEE Trans. On* Systems, 1975 pp.390-394.

[2] B.William. K.Thompson, M. Mutch and B.Valdis, "Dynamic Occlusion Analysis in Optical Flow Fields," *in IEEE Trans. on Pattern Analysis and Machine Intelligence*, in 1985,pp.374-383.

[3] Brian G. Schunck, "Image Flow Segmentation and Estimation by Constraint Line Clustering," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1989, pp.1010-1027.

[4] R. C. Jain, "Segmentation of frame sequences obtained by a moving observer," in *IEEE Trans. PAMI,* September 1984, pp.624–629.

[5] E.P.Ong, B.J.Tye, W.S. Lin, "A Fast Video Object Segmentation Scheme for MPEG-4 Video Coding ," Nanyang Technological University, Innovation Centre.

[6] X.Zhong, X.Huang, J.Wang, Z.He, "Video object segmentation based on HOS and multi-resolution watershed" in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on, Oct. 2004*, pp.274- 277.

[7] B.P.L. Lo and S.A. Velastin, “Automatic congestion detection system for underground platforms,” in *Proc. of* 2001 *International.Symp. On Intell. Multimedia, Video and Speech Processing, 2000*, pp. 158-161.

[8] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time Tracking of the Human Body,” in *IEEE Trans. on Patt. Anal. and Machine Intell., 1997, vol. 19, no. 7,* pp. 780-785.

[9] C. Stauffer, W.E.L. Grimson, “Adaptive background mixture models for real-time tracking”, *in Proc. of CVPR 1999*, pp. 246-252.

[10] D. Farin, P. H. N. de With, and W. Effelsberg, "Video-object segmentation using multi-sprite background subtraction, " in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2004, pp. 343 - 346 Vol.1.

[11] D. Farin, P. H. N. de With, and W. Effelsberg, "Video-object segmentation using multi-sprite background subtraction, " in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2004, pp. 343 - 346 Vol.1.

[12] G. Wolberg, S. Zokai, “Image registration for perspective deformationrecovery,” in *SPIE Conference on Automatic Target Recognition,* Orlando, Florida, USA, April 2000, pp.12.

[13] R.Cucchiara, C.Grana, A.Prati, M.Piccardi, “Detecting Objects, Shadows and Ghosts in *Video Streams by Exploiting Color and Motion Information,” in ICIAP 2001*, pp.360-365.

[14] ITU-R, “Methodology for Subjective Assessment of the Quality of Television Pictures,” in *Recommendation BT.500-7*, 1995.

[15] ITU-T, “Subjective Video Quality Assessment Methods for Multimedia Applications ,” in *Recommendation P.910*, August, 1996.

[16] H.G.Barrwo, J.M.Tenenbaum, R.C.Bolles and H. C. Wolf, “Parametric correspondence and chamfer matching: Two new techniques for image matching,” *in Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* Cambridge, MA, 1997, pp. 659-663.