Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review and Comparison of Well-Known Methods for Object Detection and Tracking in Videos

Ankita Rawat
M.Tech,Dept of CSE,
Graphic Era University, Dehradun, India

Anuj Saxena
CEO, Institute De Informatica
Dehradun, India

Abstract: Moving object detection and tracking are main initial steps in object recognition, context analysis and index processes for visual surveillance system. IT is a big challenge for researchers to make a decision on which detection & tracking algorithm is more suitable. There is a variety of object D&T algorithms (i.e. methods) and publications on their performance comparison. This paper provides a systematic review of these algorithms and performance measures.

Keywords- object detection, object tracking, D&T algorithms

I. INTRODUCTION

Detecting and tracking moving objects are frequently used as computer vision applications, like video surveillance, validate systems, user interfaces by gesture, robotics. Detection and tracking of vehicles is a huge problem in ITS (intelligent transportation system). For this they need first detecting the vehicle and segmenting them from the video images, and then track them from across different frames. Strong detection and tracking of vehicles on the road based on video is a huge problem. Roads are self-motivated environment, with the background changes. There is high changeability in the appearance of vehicles. Video vehicle detection is a procedure of detection the existence or nonexistence of a vehicle in the sequences. The result of detection is used as initialization procedure for tracking. After vehicle detection, ITS will perform the task of vehicle tracking. Vehicle tracking is a process that tracks the route of vehicle over time by locating its location in every frame of the video sequences.

II. METHODOLOGY

Object tracking, in general, is a challenging problem. Difficulties in tracking objects can arise due to abrupt object motion, changing appearance patterns of the object and the scene, nonrigid object structures, object-to-object occlusions, and camera motion. In its simplest form, tracking can be defined as the problem of estimating the trajectory of an object in the image plane as it moves around the scene. In other words, a tracker assigns consistent labels to the tracked objects in different frames of a video.

One can simplify tracking by imposing constraints on the motion and/or appearance of objects. For example, almost all tracking algorithms assume that the object motion is smooth with no abrupt changes. One can further constrain the object motion to be of constant velocity or constant acceleration based on a priori information. Prior knowledge about the number and the size of objects, or the object appearance and shape, can also be used to simplify the problem.

III. OBJECT DETECTION AND TRACKING METHODS

A Camera is the basic sensing element, and it is the first step for a good visual surveillance system's object D&T process. Digital signal and image processing are starter levels of digital video processing. In digital video processing, the object detection process affects object tracking and classification processes, as well. Manual Object D&T is a tedious task. For this reason, experts in computer vision research area studied for a long time semiautomatic or automatic D&T techniques. These techniques often involve maintaining a model.

A video's hierarchical structure units are scene, shot, and frame. Frame is the lowest level in the hierarchical structure. Video-content analysis, content-based video browsing and retrieval use these units. In that situation, video applications must partition a video sequence into shots.

3.1 Background Subtraction Method

In video processing applications, variants of the background subtraction (BS) method are broadly used for the detection of moving objects in video sequences. The BS's speed in locating the moving objects makes it attractive for the users. Unfortunately, a simple inter-frame difference with global threshold reveals itself as being sensitive to phenomena of the basic assumptions of BS. These assumptions are based on a firmly fixed camera with a static noise-free background. Real-life systems have camera jitters, illumination changes and etc. [1]. In object detection, usually a scene can be represented by a model called background model.

Also, the related algorithm (or method) finds the deviations from the background model for each incoming frame (i.e. frame differencing). A pixel-level background model is generated and maintained to keep track of the time-evolving

background. A moving object can be defined as any significant change in an image region compared to the background model. Intra-regions pixels' undergoing changes are marked for further processing. Usually, a connected component algorithm is applied to obtain connected regions corresponding to the objects. Background maintenance is the essential part, which may affect the performance of BS in the time-varying situations. This process is referred to as the BS (as mentioned in the survey study of Yilmaz et al.) [1].The methods of basic BS employ usually a single reference image corresponding to an empty scene as the background model. This kind of simple model was not suitable for real world's much complex surveillance systems.

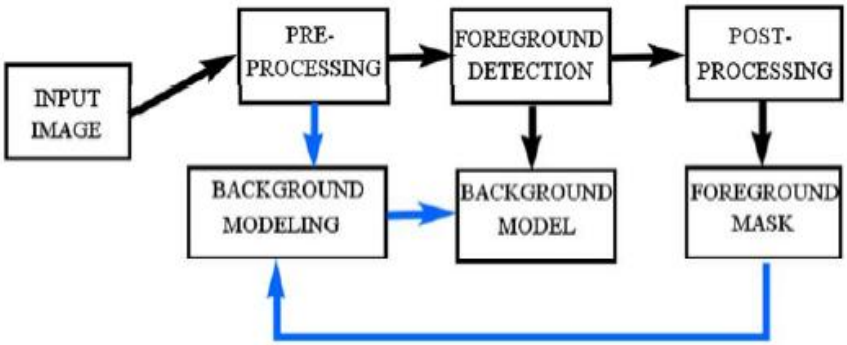

Figure 1. BS based object D&T system's architecture.

In the study by Benezeth et al. [2] common BS techniques were reviewed. In principle, according to Benezeth et al., these techniques assume the hypothesis that the observed video sequence I is made of a fixed background B, in front of which moving objects are observed. With the assumption that a moving object at time t has a color (or a color distribution, or any other desired feature) different from the one observed in B, the principle of BS methods can be summarized by the following formulation [2]:

$$\psi_t(s) = \begin{cases} 1 & if\ d(I_{s,t}, B_s) > \tau \\ 0 & otherwise \end{cases} \qquad (1)$$

where ψ_t is the motion label field at time t and a function of s(x,y) spatial location (also called motion mask), d is a distance between $I_{s,t}$ the video frame at time t at pixel s and B_s the background at pixel s; τ is a threshold. The main difference among most of the BS methods is how well B is modeled and which distance metric d is being used (e.g. Euclidean, Mahalanobis or Manhattan, etc.) [2].

In the literature, there are various BS techniques, such as basic motion detection, one gaussian, gaussian mixture model, kernel density estimation, inter-frame minimum, maximum difference and etc. [2]. In a BS algorithm, the four main steps are preprocessing, background modeling, foreground detection, and data validation, as declared in Cheung and Kamath's study [3]. Preprocessing step involves some simple image processing tasks. These tasks change the raw input video sequence into a format, which is used in subsequent steps. In background modeling, the new video frame can be used in calculation and updating of a background model, which provides a statistical description of the entire background scene (which may be static or dynamic). Much research has been devoted to developing a background model that is robust against environmental changes in the background, but sensitive enough to identify all moving objects of interest [3]. One can classify background modeling techniques into two broad categories (non- recursive and recursive). Non-recursive techniques are frame differencing, median filter, linear predictive filter and non-parametric model. Recursive techniques are the approximated median filter, Kalman filter and Mixture of Gaussians (MoG). Details of the aforementioned techniques can be found at Yilmaz et al., Benezeth et al., Cheung and Kamath, and Fuentes and Velastin studies [1-4]. In the foreground detection step, pixels in the video frame, which are not explained enough by the background model [5], are defined as a binary candidate foreground mask. The most important limitation of background subtraction is the requirement of fixed (i.e. stationary) cameras. Camera motion usually distorts the background models and causes false or partial object detection.

All the aforementioned techniques use a single image as their background models, except the non-parametric model and the MoG model [1, 2]. There are some foreground detection approaches and the most commonly-used one is to check whether the input pixel is significantly different from the corresponding background estimate:

$$\left|I_t(x,y) - B_t(x,y)\right| > \tau \qquad (2)$$

Where $I_t(x, y)$ and are used to denote the luminance pixel intensity and its background estimate at spatial location (x, y) and time t. In addition, another popular foreground detection scheme is to apply a threshold based on the normalized statistics :

$$\frac{\left|I_t(x,y) - B_t(x,y) - \mu_d\right|}{\sigma_d} > \tau_s \qquad (3)$$

where μ_d and σ_d are the mean and the standard deviation of $I_t(x, y)$-$B_t(x,y)$ for all spatial locations (x,y). In these formulations, τ and τ_s are used to denote the foreground threshold and the statistical foreground threshold, which are experimentally determined by the most of foreground detection schemes. Ideally, the threshold should be a function of the spatial location(x, y). For example, the threshold should be smaller for regions with low contrast. Sometimes, this is an advantageous situation for object detection in low contrast scenes. Fuentes and Velastin [4] proposed one possible modification for threshold determination:

$$\frac{\left|I_t(x,y)-B_t(x,y)\right|}{B_t(x,y)} > \tau_c \qquad (4)$$

where τ_c is used to denote the contrast threshold. The contrast enhancement of bright images, such as an outdoor scene under heavy fog or spot (e.g. sun spot or other flash light source spot) is not possible with this technique. In data validation step, method reviews this candidate mask and eliminates those pixels that do not correspond to actual moving objects, and outputs the final foreground mask. Real-time processing is still feasible as computationally-intensive vision algorithms are applied only on the small number of candidate foreground pixels. There are some other alternate approaches for background subtraction, which are to represent the intensity variations of a pixel in an image sequence as discrete states corresponding to the events in the environment. In some studies [5, 6], Hidden Markov Models (HMM) are used to classify small blocks of an image as belonging to one of three different states: background state, foreground state and shadow state. HMM is successful for certain events, which are hard to model correctly using unsupervised background modeling approaches, can be learned using training samples.

3.2**. Active Contour Model**

An another image segmentation approach is active contour models (ACMs), which are in scope of edge-based segmentation and used on object tracking process, as well. A snake is an energy minimizing spline, which is a kind of active contour model. The snake's energy is based on its shape and location within the image. Desired image properties are usually relevant to the local minima of this energy. ACMs, which was suggested by Kaas et. al. [7] in 1987 is also known as Snake method.

A snake can be considered as a group of control points (or snaxel) connected to each other and can easily be deformed under applied force. According to the study of Dagher and Tom [8], the situation, in which a snake works the most abundant is the situation where the points are at the adequate distance and the situation, in which the initial position's coordinates are controlled. Total energy of a snake is defined as [8] in equation 9.

$$E_{Snake} = \frac{1}{2}\int_S \left[\alpha(s)\left|v_S\right|^2 + \beta(s)\left|v_{SS}\right|^2\right] + \gamma E_{External}(v(s))ds \qquad (9)$$

where;

$$v_s = \frac{dv(s)}{ds} \qquad (10)$$

and

$$v_{ss} = \frac{d^2v(s)}{ds^2} \qquad (11)$$

Existence of a spline, on which E energy is constant, is a problem on which Euler-Lagrange differential equation can be practiced. According to this, it turns to:

$$\alpha(s)v_{SS} - \beta(s)v_{SSSS} - \nabla E_{External} = 0 \qquad (12)$$

The third term in equation 12 has been normalized by γ, the external energy weight. By solving equation 12, the final contour, which provides the minimization of E_{snake}, is obtained. It is also possible to interpret the equation in the equation 12 as an equation of balance force [8].According to this

$$F_{Internal} + F_{External} = 0 \qquad (13)$$

where;

$$F_{Internal} = \alpha(s)v_{SS} - \beta(s)v_{SSSS} \qquad (14)$$

and

$$F_{Internal} = \nabla E_{External} \qquad (15)$$

The ssss notion denotes fourth-order partial derivative. Using the final solution, gradient forces F_x and F_y are applied on the point (x,y) in x and y directions, respectively [7, 8]. There are some snake-based methods and implementations in video object segmentation and tracking. In Gouet-Brunet and Lameyre's study [9], a snake model's control points (snaxels) are taken as global descriptors describing the lobal shape of objects(eg. Interest point of objects of interest).

3.3.**Region-based Method**

When a moving object is segmented, a region of pixels assigned to the object is available. This region can be tracked using approaches like cross-correlation. The location of the region in the next frame is to be determined. A moving object usually corresponds to one or several tracked regions. Combination of several regions to one object is then performed at a higher level of abstraction.

Several techniques are available for modelling and tracking image regions. The regions are often modelled using a probability density distribution of their colour. This distribution can be described using a colour histogram [10] , or a mixture of Gaussian kernels [11]. Instead of using one 3D probability density distribution, separate distributions for each of the colours can be used.

Probability density distributions of the colour are relatively invariant to changes in object orientation, scale, partial occlusion, viewing position and object deformation. This makes them particularly interesting for tracking nonrigid objects such as humans. However, the distributions capture Region-Based Moving Object Detection and Tracking only the colours in an image and do not include any spatial correlation information. Therefore, they have limited discriminative power. A colour correlogram, on the other hand, is a cooccurrence matrix that gives the probability that a pixel at a distance d from a given pixel of colour is of colour . This way spatial information in the form of distance to pixels of a certain colour is introduced [12].

Other approaches taking spatial information into account are using many small regions and using the time average per-pixel colour . Instead of choosing one colour space, automatic selection of most discriminative features can be used . This adapts the colour space used by comparing the specific tracked region with the local background, leading to more precise object segmentation. However, when pixels are misclassified and consequently used for updating the wrong model, this solution will become unstable.

Considering the low number of pixels in each tracked region, histograms will become quite sparse. On the other hand, it also is not easy to estimate the parameters of a Gaussian mixture model from only a few data points, especially when also the number of kernels is unknown. From the point of view of computational complexity, the use of a histogram approach is most affordable because template matching can be performed very efficiently.

Considering that a probability density function is available with these techniques, it is unfortunate that object segmentation is often based on a static threshold. Calculated probabilities could be used in a probabilistic foreground/background classification algorithm.

Moving objects can also be modelled using a fixed or parameterized shape, like in the mean-shift approach [13], and the particle filter [14]). Disadvantage of such techniques is that they are unable to describe an arbitrary shape, changing between subsequent frames.

3.4**. The Feature-Based Approach**

The method of finding image displacements which is easiest to understand is the feature-based approach. This finds features (for example, image edges, corners, and other structures well localized in two dimensions) and tracks these as they move from frame to frame. This involves two stages. Firstly, the features are found in two or more consecutive images. The act of feature extraction, if done well, will both reduce the amount of information to be processed (and so reduce the workload), and also go some way towards obtaining a higher level of understanding of the scene, by its very nature of eliminating the unimportant parts. Secondly, these features are matched between the frames. In the simplest and commonest case, two frames are used and two sets of features are matched to give a single set of motion vectors. Alternatively, the features in one frame can be used as seed points at which to use other methods (for example, gradient-based methods to find the flow.

The two stages of feature-based flow estimation each have their own problems. The feature detection stage requires features to be located accurately and reliably. This has proved to be a non-trivial task, and much work has been carried out on feature detectors [15]. If a human is shown, instead of the original image sequence, a sequence of the detected features (drawn onto an *empty* image), then a smoothly moving set of features should be observable, with little feature flicker. The feature matching stage has the well known correspondence problem of ambiguous potential matches occurring; unless image displacement is known to be smaller than the distance between features, some method must be found to choose between different potential matches.

IV. **RESULT**

Some applications of well-known methods and algorithms are given in the literature. We implemented background subtraction algorithm for object detection and tracking. Fig[2] shows BS object tracking.

Fig[2]. Object tarcking

V. COMPARISON

In some surveys and studies the issue of evaluating the performance of video surveillance systems (system's total performance) is becoming more and more important.

In this paper, we review a number of well-known methods and their algorithm's comparison [16] of moving object D&T.

5.1**. Background subtraction method**: This method is one of the widely used methods to detect moving vehicle regions. It subtracts the generated background image from the input image frame to detect the moving vehicle regions. This difference image is then thresholded to extract the vehicle regions. The problem with the stored background frame is that they are not adaptive to the environment changes which may create non-existent vehicle regions and also works for stationary background.

5.2. **Active contour based method:** This method represents vehicle by bounding contour of the object and dynamically update it during the tracking. The advantage of active contour tracking over region-based tracking is the reduced computational complexity. But the disadvantage of the method is their inability to accurately track the occluded vehicles and tracking need to be initialized on each vehicle separately to handle occlusion better.

5.3**. Region based method:** This method subtracts image frame containing vehicles from the background frame which is then further processed to obtain vehicle regions (blobs). Then these vehicle regions are tracked. It can work well in free flowing traffic conditions, but the disadvantage is that it has difficulty in handling shadows and occlusion.

5.4. **Feature based method:** This method extracts suitable features from the vehicle regions and these features are processed to track the vehicles correctly. The method has low complexity and also can handle occlusions well. The disadvantage is the recognition rate of vehicles using tow-dimensional image features is low,and the problem that which set of sub features belong toone object is complex.

VI. CONCLUSION

In this paper, we review a number of commonly-implemented object D&T algorithms. Novelty of this study as related to other reviews [16] is that we made comparison and evaluation for all above mentioned D&T methods are presented. As a result of comparisons, no method outperforms the other ones on each video category. More research, however, is needed to improve the robustness against the effects of the environment such as noise, illumination changes, occlusions and etc. The variety of metrics and datasets allows us to reason about the weaknesses of particular algorithms against specific challenges. The aim of this paper is to provide a better understood of performances of video surveillance systems in the literature via published measures, computational and environmental details. These details have a large impact on the comparison of given algorithms.

REFERENCES

[1] Yilmaz, A., Javed, O., Shah, M., "Object tracking: A survey", *ACM Computing Surveys*, Vol. 38, Dec. 2006.

[2] Benezeth, Y., Jodoin, P.M., Emile, B., Laurent, H., and Rosenberger, C., "Review and evaluation of commonly-implemented background subtraction algorithms", *International Conference on Pattern Recognition (ICPR 2008)*, 8-11 Dec. 2008.

[3] Cheung, S.-C., Kamath, C. "Robust techniques for background subtraction in urban traffic video" *Video Communications and Image Processing, SPIE Electronic Imaging*, January 2004.

[4] Fuentes, L., Velastin, S., "From tracking to advanced surveillance", *In Proceedings of IEEE International Confererence on Image Processing*, Barcelona,Spain,2003.

[5] Rittscher, J., Kato, J., Joga, S., and Blake, A., "A probabilistic background model for tracking", *In European Conference on Computer Vision (ECCV)*, Vol. 2, 2000.

[6] Stenger, B., Ramesh, V., Paragios, N., Coetzee, F., and Buhmann, J., "Topology free hidden markov models: Application to background modeling", *In IEEE International Conference on Computer Vision (ICCV)*, 2001.

[7] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: active contour models", *International Journal of Computer Vision*, Vol. 1, No. 4, 1988.

[8] Dagher, I., Tom, K. E., "WaterBalloons: A hybrid watershed Balloon Snake segmentation", *Image and Vision Computing*, Vol. 26, 2008.

[9] Gouet-Brunet, V., Lameyre, B., "Object recognition and segmentation in videos by connecting heterogeneous visual features", *Computer Vision and Image Understanding*, Vol. 111, *Special Issue on Intelligent Visual Surveillance (IEEE)*, 2008.

[10] Johnson I. Agbinya and David Rees,"Multi-object tracking in video", *Real-Time Imagin*, 1999.

[11] Stephen J. McKenna, Yogesh Raja, and Shaogang Gong," Object tracking using adaptive color mixture models", *In Asian Conference on Computer Vision (ACCV)*, 1998.

[12] M.B. Capellades, David S. Doermann, Daniel DeMenthon, and RamaChellappa," An appearance based approach for human and object tracking,"*In Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2003.

[13] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. "Real–time tracking of non– rigid objects using mean shift", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[14] Patrick P´ erez, Carine Hue, Jaco Vermaak, and Michel Gangnet,"Color–based probabilistic tracking", *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2002.

[15] H. Asada and M. Brady", The curvature primal sketch", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1986.

[16] Zhiming Qian, Hongxing Shi , Jiakuan Yang and Lianxin Duan, "Video-based multiclass vehicle detection and tracking", *IJCSI International Journal of Computer Science Issues*, January 2013.

