SIMULTANEOUS BACKGROUND AND FOREGROUND MODELING FOR TRACKING IN SURVEILLANCE VIDEO

Jie Shao, Shaohua Kevin Zhou and Rama Chellappa

Center for Automation Research and Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
{shaojie,shaohua,rama}@cfar.umd.edu

ABSTRACT

We present a stochastic tracking algorithm for surveillance video where targets are dim and at low resolution. The algorithm builds motion models for both background and foreground by integrating motion and intensity information. Some other merits of the algorithm include adaptive selection of feature points for scene description and defining proper cost functions for displacement estimation. The experimental results show tracking robustness and precision in a challenging video sequences.

1. INTRODUCTION

Tracking small dim objects in video is an important and challenging research topic. Many algorithms are available for precisely tracking an object in consecutive frames, [4], [3]. Experiments show that these algorithms work very well on large objects, but less so for small objects, especially in surveillance applications, in which there may only be tens of pixels on the target. Background subtraction [1] is a much studied technique for object tracking. Its principle is to extract the moving foreground object by subtracting a long-time-average background from the current frame. Typically, pixels belonging to the objects are expected to be different in intensity, chromaticity values and motion compared to the background pixels. A static background and some prior knowledge of the original background image are assumed. Such requirements limit the employment of background-subtraction algorithm for object tracking.

We present a new statistical method for tracking objects in a surveillance video, which uses a time series state space model parameterized by a tracking motion vector, denoted by θ. In order to overcome challenges due to differences between the intensities of background and foreground object not being significant or insufficient target pixels, we integrate the intensity information with motion information. Motion information helps to discriminate moving objects and relatively still background. Therefore in the proposed algorithm time differencing images are used as measurement of observation to estimate the motion state of the model. To compensate for the camera motion, the original images are first stabilized using a subset of parameters included in the motion vector θ. Hence, the motion vector θ for the entire system consists of two parts:

$$\theta = \{\theta^F; \theta^B\} = \{dx^F, dy^F; \alpha^B, dx^B, dy^B, s^B\} \quad (1)$$

Partially funded by the DARPA VIVID program through a subcontract from SRI international.

where θ^B represents background parameters, θ^F represents foreground parameters. Image rotation angle α, displacement dx^B, dy^B and scale s are four parameters in θ^B describing the background changes caused by the moving camera, while foreground object displacements dx^F, dy^F in θ^F represent the movement of an object occupying a small number of pixels.

A particle filter is then developed to provide a numerical approximation to the posterior distribution of the motion vector at time *t* given the observation up to *t*, i.e.,$p(\theta_t|Y_{1:t})$ where $Y_{1:t}$ are the observations up to time *t*. Since the system can simultaneously estimate both background and foreground motion parameters, the implementation is very efficient for: 1) segmenting background and foreground objects; 2) obtaining motion information from background motion vector θ^B; and 3) tracking the foreground target based on foreground motion vector θ^F. The proposed algorithm uses intensity information together with motion information. It can track moving targets at a very small scale.

The rest of the paper is organized as follows. After a theoretical discussion of the proposed tracking algorithm in Sec. 2, we describe the detailed implementation. Experimental results and discussions are presented in Sec. 3, followed by Sec. 4, which concludes the paper.

2. BACKGROUND-FOREGROUND TRACKING USING PARTICLE FILTER

2.1. System Description on Particle Filter Model

The particle filter was originally proposed in [2] in the signal processing literature and has been used to solve many vision tasks [8] [6]. Given the state transition model characterized by the state transition probability $p(\theta_t|\theta_{t-1})$ and the observation model characterized by the likelihood function $p(Y_t|\theta_t)$, the particle filter is used to approximate the posterior distribution $p(\theta_t|Y_{1:t})$ by a set of weighted particles $S_t = \{\theta_t^{(j)}, \omega_t^{(j)}\}_{j=1}^J$ with $\Sigma_{j=1}^J \omega_t^{(j)} = 1$.

As mentioned in section 1, we define the state motion vector θ by (1). Accordingly, the posterior distribution becomes a joint distribution of the background and the foreground, which is $p(\theta_t^B; \theta_t^F|Y_{1:t})$. Often, we need to do background stabilization when estimating the foreground motion. That means in practice, θ_t^F is estimated conditioned on knowing θ_t^B, so that the relation between background and foreground distributions can be written as

$$p(\theta_t^B, \theta_t^F|Y_{1:t}) = p(\theta_t^B|Y_{1:t})p(\theta_t^F|\theta_t^B, Y_{1:t}) \quad (2)$$

Using this assumption, time recursiveness and Markov property,

we can easily derive

$$p(\theta^B_{1:t}, \theta^F_{1:t}|Y_{1:t}) = p(\theta^B_{1:t}|Y_{1:t})p(\theta^F_{1:t}|\theta^B_{1:t}, Y_{1:t}) \quad (3)$$

$$p(\theta^B_{1:t}|Y_{1:t}) = p(\theta^B_{1:t-1}|Y_{1:t-1})\frac{p(Y_t|\theta^B_t)p(\theta^B_t|\theta^B_{t-1})}{p(Y_t|Y_{1:t-1})} \quad (4)$$

$$p(\theta^F_{1:t}|\theta^B_{1:t}, Y_{1:t}) = p(\theta^F_{1:t-1}|\theta^B_{1:t-1}, Y_{1:t-1})\frac{p(Y_t, \theta^B_t|\theta^F_t)p(\theta^F_t|\theta^F_{t-1})}{p(Y_t, \theta^B_t|Y_{1:t-1}, \theta^B_{1:t-1})} \quad (5)$$

Hence, the posteriors probability is recursively determined by the likelihood functions $p(Y_t|\theta^B_t)$, $p(Y_t, \theta^B_t|\theta^F_t)$ and state transition probabilities $p(\theta^B_t|\theta^B_{t-1})$, $p(\theta^B_t|\theta^B_{t-1})$. It also indicates a feasible strategy for estimating the likelihood functions: We estimate $p(Y_t|\theta^B_t)$ first, then based on the approximation $\hat{\theta}^B_t$, we estimate $p(Y_t, \hat{\theta}^B_t|\theta^F_t)$, instead of $p(Y_t, \theta^B_t|\theta^F_t)$. It is valid as $\hat{\theta}^B_t$ is an optimum estimate of θ^B_t when the entire parameter estimation problem is regarded as a two-step optimization problem.

2.2. Background Motion Vector Estimation

To obtain the difference images, the simplest way is to take consecutive frames and directly find the absolute differences. This is based on the assumptions that the camera is stationary and the image is noise-free. Obviously such assumptions are rarely valid in practice. Stabilization has to be used to compensate for the inter-frame differences caused by the camera motion.

Let $\mathbf{X} = (x, y)^T$, $d\mathbf{X}_t = (dx, dy)^T$. We then have:

$$\mathbf{X}_t = s \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \mathbf{X}_{t-1} + d\mathbf{X}_t \quad (6)$$

where s is the scaling factor, α is the rotation angle between the two frames, $d\mathbf{X}_t$ is the translation measured in the image coordinates at time t. According to the transform equation, four parameters $(\alpha^B_t, dx^B_t, dy^B_t, s^B_t)$ are used to describe the motion of the background between frame $t-1$ and t, where α^B_t is the rotation angle, dx^B_t and dy^B_t represent translation parameters, and s^B_t is the scale factor. These parameters characterize the motion of the camera. The state transition is then approximated using a first-order Markov chain and a mixture Gaussian noise model. One component of noise is with zero-mean Gaussian distribution, denoted as μ_t, accounting for sensor noise, digitization noise etc; the other is a non-zero-mean Gaussion distribution, denoted as ν_t, due to camera motion. The equation is expressed as:

$$\theta^B_t = \theta^B_{t-1} + \nu_t + \mu_t = \widetilde{\theta}^B_t + \mu_t \quad (7)$$

Where $\widetilde{\theta}^B_t$ is the initial approximate derived based on the stabilization algorithm proposed in [7].

2.3. Foreground Motion Vector Estimation

Once we estimate the background motion parameters, we generate the stabilized difference images. Let $\Delta_t = I_t - T_{\hat{\theta}^B_t} I_{t-1}$, where Δ_t is the difference image and I_t and I_{t-1} are the original frames, $T_{\hat{\theta}^B_t}$ is the stabilizing function with $\hat{\theta}^B_t$. The advantage of using Δ_t instead of I_t is: relative to the background, most of the targets in a surveillance video are in motion. Motion information turns out to be a decisive measurement to identify the foreground object from the background, especially when the intensities of background and foreground are not that different. Two extreme examples are illustrated in the first row of Fig 1, where it is visually difficult to separate the background from the foreground. But if we use the difference images shown in the second row, the targets are easily picked. The third row shows the edge maps of the difference images in which the targets are more easily detected. To analyze

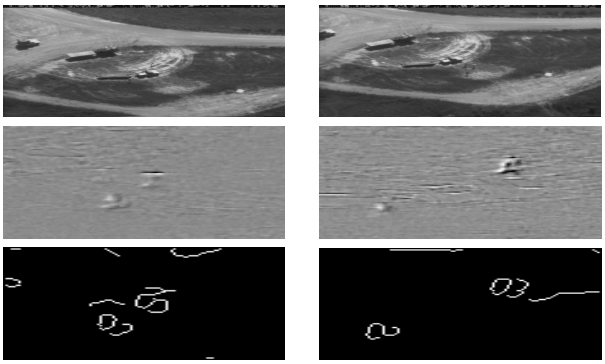


Fig. 1. Examples of people with low contrast imagery passing through an open field. (upper) original frames; (middle) part of Δ images containing targets; (lower) part of the edge maps containing targets.

the motion of a foreground target, we process the difference images, instead of the original frames. For delineating the inside and the outside regions of the moving object, the edge gradient information of the Δ images is uses. The edge image E is generated as $E = \Delta \otimes DoG$, where DoG is a 2D derivative of Gaussian filter. We assume that the motion of one feature pixel on the target can describe the motion of the entire target. A support region is applied to each pixel to collect enough edge gradient measurement for computing the cost function. In addition, we assume that dx^F and dy^F are independent, and that their joint distribution is uniformly distributed in a specified 2D space $\mathcal{D}_2$. The function collects the match measurement in a pre-defined rectangle region $\mathbf{W_X}$ with $\mathbf{X}$ as its center.

The idea behind the foreground displacements estimation is: A particular pixel in Δ_{t-1} image will move by a distance in Δ_t image due to the motion continuousness. This is estimated by matching a region between the successive Δ images using a cost function $D(\mathbf{X}; d)$ where d denotes the linear translation $d = (d_x, d_y)$. The cost function and the likelihood function are defined respectively as:

$$D(\mathbf{X}; d) = \sum_{\mathbf{Y} \in \mathbf{W_X}} \frac{|E_t(\mathbf{Y}) - E_{t-1}(\mathbf{Y} + d)|}{|\mathbf{W_X}|} \quad (8)$$

$$p(Y_t, \theta^B_t|\theta^F_t) \propto \exp\left(-\frac{D(\mathbf{X}; d)}{2\sigma_1^2}\right) \quad (9)$$

where $|\mathbf{W_X}|$ is the number of pixels in the window $\mathbf{W_X}$, $\mathbf{X}$ represents a pixel position in the image $(x, y)^T$.

2.4. Implementation

Assuming that the start time is t_0, the algorithm is:

Step.1. Initialization: The initial location $\mathbf{X}_0 = (x_0, y_0)^T$ of object, the initial Δ_0, E_0 and an initial sample set $S_0 = \{\theta^{(m)}_0, 1\}^M_{m=1}$ are set;

Step.2. Iteration: This step realizes the frame-by-frame tracking. At each time instant, t=1,2,...:

Step.2.1 Background motion parameters θ_t^B estimation:

Step.2.1.1 Select feature points, approximate $\widehat{\theta}_t^B$ as an initial motion vector using the method proposed in [7];

Step.2.1.2 For m=1,2,...,M_1, where M_1 is the number of samples for background motion estimation: Draw noise sample $\mu_t^{(m)}$, approximate $\widetilde{\theta}_t^{B(m)}$ using (7), compute the cost function and the likelihood function, defined by

$$p(Y_t|\theta_t^B) \propto \exp\left(-\frac{\|\breve{\mathbf{I}}_t - T_{\theta_t^B}\{\breve{\mathbf{I}}_{t-1}\}\|}{2\sigma_2^2}\right) \quad (10)$$

where $\breve{\mathbf{I}}_t$ represents the original frame image with the foreground region cut out.

Step.2.1.3 Let $\hat{\theta}_t^B = \arg\max_{\theta_t^{B(m)}, m\in(1,\ldots,M_1)} p(Y_t|\theta_t^B)$, compute Δ_t with parameters in $\hat{\theta}_t^B$;

Step.2.2 Foreground motion parameters θ_t^F estimation:

Step.2.2.1 Compute the edge gradient image E_t;

Step.2.2.2 For m=1,2,...,M_2, where M_2 is the number of samples for foreground motion: Draw the displacement sample $d^{(m)} = (dx^{(m)}, dy^{(m)})$; Compute the cost function $D(\mathbf{X}_{t-1}, d^{(m)})$ and the likelihood function $p(Y_t, \hat{\theta}_t^B|\theta_t^F)$ according to (8) and (9);

Step.2.2.3 Let $\hat{\theta}_t^F = \arg\max_{\theta_t^{F(m)}, m\in(1,\ldots,M_2)} p(Y_t, \hat{\theta}_t^B|\theta_t^F)$;

Step.2.2.4 Update the current target location:

$$\begin{aligned}\hat{\mathbf{X}}_t &= \hat{\mathbf{X}}_{t-1} + d\hat{\mathbf{X}}_t^F + d\hat{\mathbf{X}}_t^B \\ &= (x_{t-1}, y_{t-1})^T + (dx_t^F, dy_t^F)^T + (dx_t^B, dy_t^B)^T \quad (11)\end{aligned}$$

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experiment Results

We present experimental results using an outdoor surveillance video sequence of moving people. The humans are very small, occupying only 20-40 pixels. The tracking results are shown in Fig 2. In these images, white boxes indicate the locations of the targets. In the experiments, two hundred corresponding feature points are selected using the optical flow method described in [5] to approximate the initial background motion parameters in step.2.1.1. The number of background noise samples is 150, the number of foreground displacement samples is 25 with $[-2:2, -2:2]$ as the distribution space $\mathcal{D}_2$, the region window $\mathbf{W}_\mathbf{X}$ is 5×5 , and the size of derivative Gaussian filter is 10×10.

3.2. Discussions

The success of our algorithm is due to the following reasons: 1) The foreground motion estimation is mainly derived from motion information, which overcomes the limitation of estimating from intensity-based method. The latter usually gives poor performance in surveillance videos. 2) With a derivative of Gaussian filter, the edge information is used to improve the robustness. Without applying the DoG filter, when target enters the shadow area, tracking fails due to environment disturbance and image noise. 3) We simultaneously track both the background and the foreground motions. Besides the fact that fixing the background parameters helps to estimate the foreground parameters, this makes it possible to exclude foreground pixels from the feature point set, which is used for estimating the background motion parameters. Thus the estimation error of the background motion parameters is lowered. 4) A stabilization algorithm is used to compensate for the error caused by moving camera. 5) The cost function we used to evaluate the displacement estimation is efficient for both non-rigid and rigid object. 6) The strategy of reselecting the feature point set at each time instant gives the system an adaptive property, which guarantees that the feature points being used for parameter approximation are always "good". As a result, this again improves the precision of the estimation of background parameters.

Among those key elements contributing to the success of the tracking, two crucial factors are: First, compared with the appearance information, the motion information is more efficient for distinguishing a moving object from the background when they are in a low-contrast situation. Second, the idea of simultaneous background/foreground modeling is effective for designing a practical tracker since target tracking can be regarded as a local discrimination problem with two classes: the foreground and the background. Tracking success or failure depends primarily on how distinguishable an object is from its surroundings. The background information naturally provides cues for delineating the foreground object from its surrounds. The motion information of the foreground object also reflects changes in the background.

4. CONCLUSION

We have presented a surveillance tracking algorithm that 1) simultaneously tracks both the background motion and the foreground motion; 2) integrates the motion and the appearance information. Using the particle filter, the approach builds a robust motion model over a state-space of multiple-hypotheses for a moving object with several novel characteristics. The experimental results demonstrate that the tracker is very efficient even on challenging video sequences with low-contrast and small objects. The idea of considering both the background and the foreground motion is very effective. We are now investigating its applications to several problems such as vehicle tracking and object classification.

5. REFERENCES

[1] T. H. Chalidabhongse, K. Kim, D. Harwood, and L. Davis. A perturbation method for evaluating background subtraction algorithms. *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Nice, France, Oct 2003*.

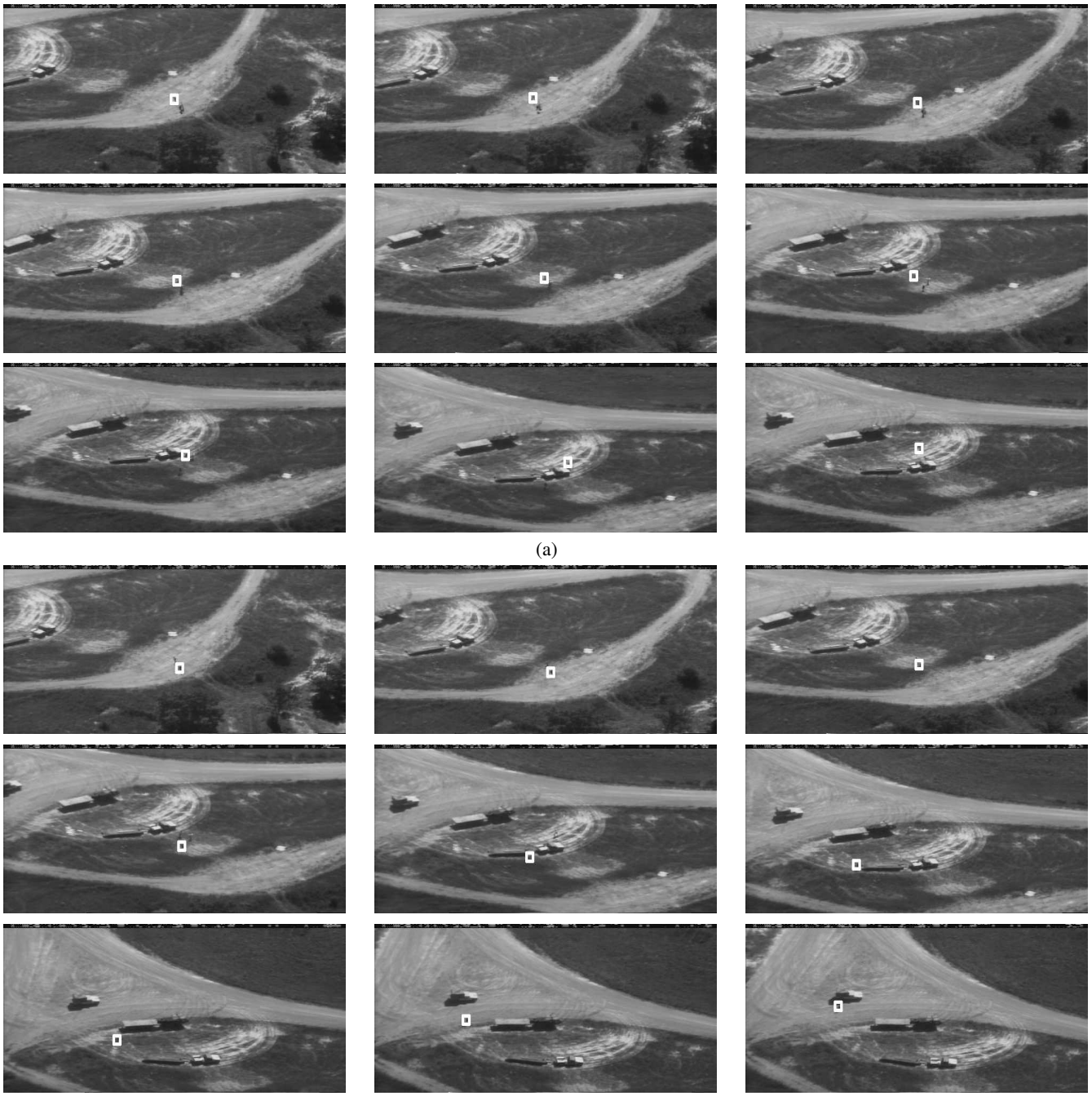
[2] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113, 1993.

[3] M. Isard and A. Blake. Contour tracking by stochastec propagation of conditional density. *Proc. of European Conference on Computer Vision, Cambridge, UK, April, 1996*, pages 343–356, 1996.

[4] T. Jebara and A. Pentland. Parameterized structure from motion for 3d adaptive feedback tracking of faces. *Proc. of IEEE Computer Society Conf. on Computer Vision Pattern Recognition, Puerto Rico, June, 1997*, pages 144–150, 1997.

[5] B. Lucas and T. Kanade. An iterative image registation technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[6] G. Qian and R. Chellappa. Structure from motion using sequential monte carlo methods. *Proc. of Intl. Conf. on Computer Vision, Vancouver, Canada, July 2001, Accepted for intl. Jl. Computer Vision*, pages 614–621, 2001.



(a)

(b)

Fig. 2. Tracking results of different persons running in the field (the white box indicates the location of the target, the top 3 rows show one person's moving, the bottom 3 rows are showing another person).

[7] Q. Zheng and R. Chellappa. A computational vision approach to image registration. *IEEE Trans. Image Processing*, 2:311–326, 1993.

[8] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *Accepeted for IEEE Trans. Image Processing*, 2003.