TRACKING OBJECTS IN VIDEO USING MOTION AND APPEARANCE MODELS

Aswin C Sankaranarayanan, Rama Chellappa and Qinfen Zheng

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

ABSTRACT

This paper proposes a visual tracking algorithm that combines motion and appearance in a statistical framework. It is assumed that image observations are generated simultaneously from a background model and a target appearance model. This is different from conventional appearance-based tracking, that does not use motion information. The proposed algorithm attempts to maximize the likelihood ratio of the tracked region, derived from appearance and background models. Incorporation of motion in appearance based tracking provides robust tracking, even when the target violates the appearance model. We show that the proposed algorithm performs well in tracking targets efficiently over long time intervals.

1. INTRODUCTION

Many vision applications involve the use of tracking algorithms. A target detection system is used as a front-end that cues the tracker. Target motion is a popular feature used for detection, and the appearance of the target is used to maintain the track. However, knowledge of target motion extracted by the detection scheme is not always used in tracking. Incorporating such information in the tracking algorithm can make the tracker robust, especially when appearance modeling of the target fails. In this paper, we propose a framework to incorporate the background model developed by the detection algorithm into the tracker. This is achieved by modeling the observation (image) as two distinct regions, one that arises from the background model of the detection algorithm and the other from the appearance model associated with the tracker. Other than the existence of the background and appearance model, the proposed framework assumes very little about detection and tracking algorithms. Stochastic modeling for both background and target appearance are assumed to be available and that the background model is assumed to treat individual pixels independently.

Prepared through collaborative participation in the Advanced Sensors Consortium sponsored by the US Army Research Laboratory under the Collaborative Technology Alliance Program, Corporate Agreement DAAD19-01-02-0008.

1.1. Prior Work

There has been significant work that combine background and appearance/foreground models. The work by Nahi and Lopez-mora [1] is in spirit the closest to the work presented in this paper. Their work deals with estimating boundaries of objects in a single image with statistical characterization of objects and background. Use of both foreground and background characterization is central to both formulations. In [2], the video is stabilized by estimating the background motion. The gradient image obtained from frame differencing the stabilized frame sequence is used to design the tracker. By using edge information, robust tracking is achieved for rigid as well as non-rigid objects. [3] illustrates the use of temporal frame differencing and shape detectors to detect location of targets. The detected targets are used to cue an appearance based tracker, for both initialization of the tracker and re-initialization when tracking fails. [4] proposed a 3-D object tracker that combines detection and tracking using optimal shape detectors. However, it is noted that there is no appearance information used for tracking, and shape features are used in place of appearance. With the exception of [1], existing methods are tuned to individual nature of the tracking and detection algorithms and are not extensible to any tracker/detector pair. In this paper, we provide a framework that makes this possible.

The paper is organized as follows: Section 2 reviews background modeling and appearance based tracking using particle filtering. In Section 3, we discuss the proposed algorithm. Section 4 presents the results of experiments on video sequences collected by ground-based and airborne sensors.

2. BACKGROUND AND APPEARANCE MODELING

The core of the proposed algorithm involves in redefining the problem at a fundamental level. To illustrate this we first interpret background modeling in a different light.

2.1. Background Modelling

Most background algorithms can be reformulated so as to define a density function $p(Y_t|B_t)$, Y_t and B_t being the observation and the background model at time (or frame) t respectively. Under the assumption that individual pixels are independent, we can define probability density functions (pdf) at each pixel. Let $p(Y_t^i|B_t^i)$ be the density for pixel i.

$$p(Y_t|B_t) = \prod_i p(Y_t^i|B_t^i) \tag{1}$$

Each pixel is subject to a hypothesis test as to whether or not it belongs to the background model. It can be shown that the likelihood ratio test results in thresholding the density $p(Y_t^i|B_t^i)$ to determine its label.

$$p(Y_t^i|B_t^i) \begin{cases} > T, & i \in \mathcal{B} \\ < T, & i \in \mathcal{F} \end{cases} \tag{2}$$

where T is the threshold and sets $\mathcal{F}$ and $\mathcal{B}$ are foreground and background pixels independently. Note that the definition of foreground for such detection algorithms is that the pixel is *not* of the background model. There is no explicit foreground modeling done at the detection stage. The above formulation chooses sets $\mathcal{B}$ and $\mathcal{F}$ such that the cost function defined below is maximized.

$$J_T(\mathcal{B}, \mathcal{F}) = \prod_{i \in \mathcal{B}} \frac{p(Y_t^i|B_t^i)}{T} \prod_{i \in \mathcal{F}} \frac{T}{p(Y_t^i|B_t^i)} \tag{3}$$

2.2. Appearance based tracking

Appearance based tracking is usually done in a particle filtering [5, 6] framework. The goal is to determine an unknown state θ of a system from a collection of observations $Y_{1:t}$. A state space model is employed to perform this estimation. This involves defining the state-transition model defining the state evolution and the observation model which specifies the state-observation dependence.

$$\text{State transition model: } \theta_t = f_t(\theta_{t-1}, u_t) \tag{4}$$

$$\text{Observation model: } Y_t = g_t(\theta_t, v_t) \tag{5}$$

where u_t is the system noise and v_t is the observation noise. Alternatively, the two models can also be defined by the probability density function $p(\theta_t|\theta_{t-1})$ and $p(Y_t|\theta_t)$. The two density functions specify the problem completely. Particle filtering proposes the weighted set $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^N$ as an approximation to the desired posterior density $p(\theta_t|Y_{1:t})$. This set is properly weighted for estimating functions of the form $E[I(\theta_t)|Y_{1:t}]$.

$$E[I(\theta_t|Y_{1:t})] \approx \sum_{j=1}^{N} I(\theta_t^{(j)}) w_t^{(j)} \tag{6}$$

The state estimate $\hat{\theta}_t$, can either be the minimum mean square error (MMSE) estimate,

$$\hat{\theta}_t = \theta_t^{\text{MMSE}} = E[\theta_t|Y_{1:t}] \approx \frac{1}{N} \sum_{j=1}^{N} \theta_t^{(j)} w_t^{(j)} \tag{7}$$

or the maximum likelihood estimate (MLE),

$$\hat{\theta}_t = \theta_t^{\text{MLE}} = \arg\max_{\theta_t} p(Y_t|\theta_t) \approx \theta_t^{\arg\max_j w_t^{(j)}} \tag{8}$$

Most appearance based tracker associate with θ target-parameters such as location, shape, size. Given an appearance model A_t at time t, the observation model can be written as

$$p(Y_t|\theta_t) = p(Y_t|\theta_t, A_t) \tag{9}$$

The appearance model may be available as an input or may be built online. In general, for online appearance modeling A_t is built as a function of observations.

$$A_t = A(\hat{\theta}_{1:t-1}, Y_{1:t-1}) \tag{10}$$

In equation 9, θ_t identifies a region of the image, denoted by $\mathcal{F}(\theta_t)$ and the observation model reduces to matching (with some distance metric or in a stochastic sense) $Y_t(\mathcal{F}(\theta_t))$ with A_t and the rest of the image is ignored. Mathematically, the rest of image can be assumed to come with an observation noise of infinite variance, thereby leading to any possible value with equal probability. In a sense, particle filtering maximizes the cost function defined as:

$$J(\mathcal{F}(\theta_t)) = p(Y_t(\mathcal{F}(\theta_t))|A_t) \tag{11}$$

With this background, we now look at the proposed framework.

3. PROBLEM FORMULATION

The proposed tracker segments the observation Y_t into two mutually exclusive and collectively exhaustive sets: $\mathcal{B}(\theta_t)$ the region identified by the state θ_t as the background and $\mathcal{F}(\theta_t)$ the region proposed by θ_t as the target. With this assumption, we now re-derive the the observation equation $p(Y_t|\theta_t)$.

$$\begin{aligned} p(Y_t|\theta_t) &= p(Y_t(\mathcal{B}(\theta_t)), Y_t(\mathcal{F}(\theta_t))|\theta_t) \\ &= p(Y_t(\mathcal{B}(\theta_t))|\theta_t) p(Y_t(\mathcal{F}(\theta_t))|\theta_t) \\ &= p(Y_t(\mathcal{B}(\theta_t))|B_t) p(Y_t(\mathcal{F}(\theta_t))|A_t) \end{aligned} \tag{12}$$

where A_t is the appearance model and B_t is the background model. Equation (12) incorporates information from the appearance and background models. Under the assumption that the background is composed of independent pixels,

$$p(Y_t(\mathcal{B}(\theta_t))|B_t) = \frac{p(Y_t|B_t)}{p(Y_t(\mathcal{F}(\theta_t))|B_t)} = \frac{p(Y_t|B_t)}{p(Y_t(\mathcal{F}(\theta_t))|B_t)}$$

1. Given particle set $\{\theta_{t-1}^{(j)}, w_{t-1}^{(j)}\}_{j=1}^{N} \sim p(\theta_{t-1}|Y_{1:t-1})$.
2. Propose new particles $\{\theta_t^{(j)}\}_{j=1}^{N}$ such that
$$\theta_t^{(j)} = g(\cdot|\theta_{t-1}^{(j)}, Y_t), \qquad j = 1, \ldots, N$$
3. Get unnormalized weight $\tilde{w}_t^{(j)}$ for $\theta_t^{(j)}$.
$$\tilde{w}_t^{(j)} = w_{t-1}^{(j)} \frac{p(Y_t(\mathcal{F}(\theta_t^{(j)}))|A_t) p(\theta_t^{(j)}|\theta_{t-1}^{(j)})}{p(Y_t(\mathcal{F}(\theta_t^{(j)}))|B_t) g(\theta_t^{(j)}|\theta_{t-1}^{(j)}, Y_t)}$$
4. Normalize weights:
$$w_t^{(j)} = \frac{\tilde{w}_t^{(j)}}{\sum_{k=1}^{N} \tilde{w}_t^{(j)}}$$
5. Set $\{\theta_t^{(j)}, w_t^{(j)}\}_{j=1}^{N} \sim p(\theta_t|Y_{1:t})$.
6. Estimate MLE Estimate θ_t^{MLE}.
7. Resample particle set if necessary.

Table 1. Proposed algorithm for combining appearance and background modeling.

With this we can now derive the desired posterior $p(\theta_t|Y_{1:t})$.

$$p(\theta_t|Y_{1:t}) \propto \frac{p(Y_t(\mathcal{F}(\theta_t))|A_t) p(\theta_t|\theta_{t-1})}{p(Y_t(\mathcal{F}(\theta_t))|B_t)} p(\theta_{t-1}|Y_{1:t-1}) \tag{13}$$

Other terms that involve the proportionality are independent of the parameter to be estimated. The term of prime interest in (13) is the ratio $\frac{p(Y_t(\mathcal{F}(\theta_t))|A_t)}{p(Y_t(\mathcal{F}(\theta_t))|B_t)}$. This term is the likelihood ratio associated with a 2 class hypothesis problem on the set $Y_t(\mathcal{F}(\theta_t))$, where the *alternate* hypothesis suggests $Y_t(\mathcal{F})$ as coming from the foreground (and hence, from the appearance model) and the *null* hypothesis suggesting that $Y_t(\mathcal{F})$ is statistically from the background model B_t. Using this for our observation model and an estimate of $\hat{\theta}_t = \theta_t^{\text{MLE}}$ as defined in (8) will result in the particle with the highest likelihood ratio being selected as the estimate. Alternatively, this chooses the particle with the minimum Bayesian risk associated with the hypothesis test. Simultaneously this formulation also minimizes the cost that the tracked region comes from the appearance model and the rest of the observation comes from the background model.

For sake of completeness, the algorithm is illustrated in Table 1. Note that the general particle filtering algorithm described in Table 1, involves the use of an importance function $g(\cdot|\theta_{t-1}, Y_t)$. The importance function is crucial for precise estimation of the pdf, especially when the sample size N is small. The choice of importance function depends heavily on the problem and the complexity of the observation and the state-transition models. However, for simplicity many trackers equate the importance function to the state transition density $p(\theta_t|\theta_{t-1})$.

The rest of the paper deals with experiments to illustrate the effectiveness and robustness of the tracker under different scenarios. It is emphasized here that in our formulation we have not made any specific assumption about the nature of the tracking and detection algorithm. Thus using our method a given pair of detection and tracking algorithms can be combined.

4. EXPERIMENTS

Experimental results are presented here for video sequences containing moving targets for both stationary and moving cameras. The algorithm described in [7] was used for tracking. The tracker builds an online appearance model and models target motion as rigid and affine on the image plane. For stationary cameras, background subtraction was performed using a fixed background that contained no targets. For moving cameras, the background model is built using the color information of the pixels surrounding the target. A mixed Gaussian form is assumed for the pdf of $p(Y_t^i|B_t^i)$, the mean, variance and mixture probability is learnt from the data.

Figure (1) shows results for tracking a person when the camera is stationary. Tracking in this data set is difficult for the appearance based tracker as the target matches the background and most trackers latch onto the background after few frames. Further, the target undergoes non-rigid deformations and this does not conform to the tracker we use. However, combining motion information with appearance modeling prevents the tracker from locking on to the background.

Figure (2) show tracking results for a moving camera. Note that the robust tracking is achieved even when the number of pixels on the target is around 10 pixels. One of the main problems for appearance based tracking when the camera is moving is that the target locks onto regions in the background that have high contrast. Incorporating motion information avoids such problems. Tracking was performed efficiently for over 600 frames. Target initialization was done manually for the first frame and in frames where new targets appeared. Tracks were also terminated manually when targets left the image.

Simultaneous modeling as proposed in this paper goes beyond normal appearance based tracker. The proposed tracker maintains track even when appearance modeling fails. This is important for a tracker that build appearance models online because modeling fails when the pose of the target changes. Further, the proposed tracker provides a statistically sound technique for combining motion and appearance without enforcing stringent constraints on the individual models used to represent/extract motion and appearance information.

5. CONCLUSION

A method has been proposed in this paper that combines motion and appearance information for tracking. The proposed tracker is robust even when appearance modeling fails or when the target size is small. The tracker maximizes

Fig. 1. Tracking a person with a stationary camera. The appearance of the person matches the background at certain instants.

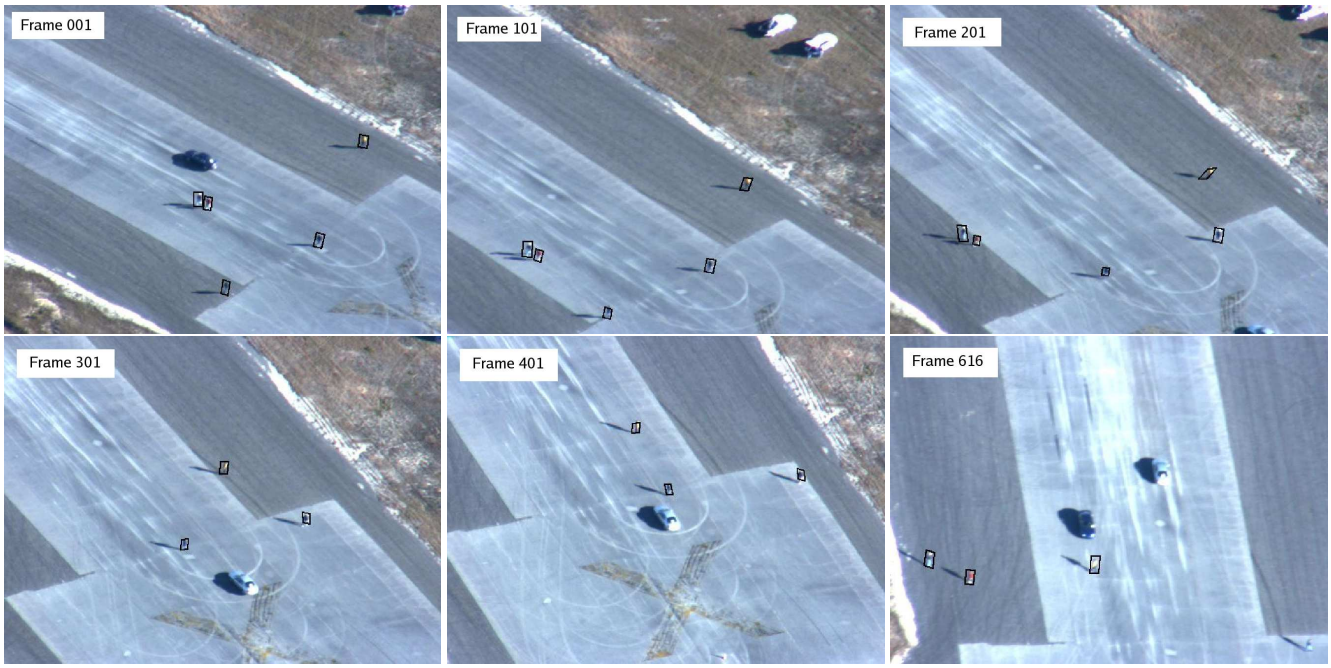

Fig. 2. Tracking multiple targets with a moving camera over 600 frames. Tracking is robust even when targets have less than 10 pixels on them.

the likelihood that the tracked region comes from the appearance model and not from the background model. Experimental results demonstrate the robustness of the tracker even when appearance modeling fails. Future works involves importance sampling that uses motion as a cue.

6. REFERENCES

[1] N. Nahi and S. Lopez-Mora, "Estimation-detection of object boundaries in noisy images," *IEEE Trans. on Automatic Control*, pp. 834–846, 1978.

[2] J. Shao, S. Zhou, and R. Chellappa, "Simultaneous background and foreground modeling for tracking in surveillance video.," in *ICIP*, 2004.

[3] J. Shao, S. Zhou, and Q. Zheng, "Robust appearance-based tracking of moving object from moving platform.," in *ICPR*, 2004, pp. 215–218.

[4] H. Moon, R. Chellappa, and A. Rosenfeld, "3d object tracking using shape-encoded particle propagation.," in *ICCV*, 2001, pp. 307–314.

[5] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *IEE Proceedings on Radar and Signal Processing*, 1993, vol. 140, pp. 107–113.

[6] A. Doucet, N. D. Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*, Springer-Verlag, New York, 2001.

[7] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle fi lters," *IEEE Trans. on Image Processing*, November 2004.