A State Vector Augmentation Technique for Incorporating Indirect Velocity Information into the Likelihood Function of the SIR Video Target Tracking Filter

Jesyca C. Fuenmayor Bello and Joseph P. Havlicek
School of Electrical and Computer Engineering
University of Oklahoma, Norman, OK, USA

***Abstract*—We consider the problem of tracking moving targets in heavily cluttered video sequences and introduce a new state vector augmentation technique to incorporate indirect velocity information into the likelihood function of the SIR particle filter. While the importance of motion information in video tracking has been well recognized, the standard SIR filter typically weights particles using a likelihood function that considers the appearance model only. Our goal is to prevent particles with poor velocity estimates from receiving large weights. The key modifications involve saving the previous values of the state variables in the state update equation and reformulating the measurement model to deliver both the current and previous observations. This leads to a straightforward calculation of likelihood across pairs of temporally adjacent frames. Our preliminary experimental results show that the proposed method is effective for avoiding track losses due to the filter locking onto structured clutter.**

***Keywords*-video tracking, particle filter, SIR, velocity estimation**

I. Introduction

Particle filters have become immensely popular for video object tracking in recent years. The sampling importance resampling (SIR) filter [1] in particular is *widely* used. First-order Markovian state dynamics are almost always assumed [2], [3]. The SIR filter's choice of the prior density function as the proposal density combined with its resampling and equal weighting of the particles at the end of every time step then imply that the particle weights are given by the likelihood function, which is straightforward to design and implement in a wide variety of practical scenarios.

It is common to assume that the target kinematics obey a constant velocity model [4], [5]. In a typical formulation, the state vector is given by, e.g.,

$$\mathbf{x}_k = [x_k\ \dot{x}_k\ y_k\ \dot{y}_k\ \gamma_k\ \dot{\gamma}_k\ \theta_k\ \dot{\theta}_k]^T, \tag{1}$$

where $[x_k\ y_k]^T$ is the target centroid, $[\dot{x}_k\ \dot{y}_k]^T$ is the velocity, γ_k and θ_k are magnification and planar rotation relative to a reference appearance model, and $\dot{\gamma}_k$ and $\dot{\theta}_k$ are the time derivatives of γ_k and θ_k. The state update and measurement models are given by [1]

$$\mathbf{x}_{k+1} = \mathbf{f}_k(\mathbf{x}_k, \mathbf{v}_k) \tag{2}$$

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k), \tag{3}$$

where $\mathbf{v}_k$ and $\mathbf{n}_k$ are mutually uncorrelated i.i.d. noises. The measurement $\mathbf{z}_k$ is typically taken as the current video frame or a set of features extracted from the current frame. Often, $\mathbf{f}_k$ and $\mathbf{h}_k$ are independent of k so that $\mathbf{f}_k(\cdot) = \mathbf{f}(\cdot)$ and $\mathbf{h}_k(\cdot) = \mathbf{h}(\cdot)$.

The SIR filter maintains a set of N_s weighted particles $\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}$ that approximate the posterior density via [1]

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i). \tag{4}$$

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-08-1-0293.

Each particle $\mathbf{x}_k^i$ is an instance of the state vector that makes a hypothesis about the true state of the target. One of the main advantages of the SIR formulation is that the weights are easy to calculate: they are given by the likelihood function

$$w_k^i \propto p(\mathbf{z}_k|\mathbf{x}_k^i), \tag{5}$$

where the proportionality constant is chosen so that the w_k^i sum to one. Elements of the state vector such as the target centroid may be estimated by taking the expected value with respect to (4) according to

$$[\widehat{x}_k\ \widehat{y}_k]^T = \sum_{i=1}^{N_s} w_k^i [x_k^i\ y_k^i]^T \tag{6}$$

or by taking (MAP) estimates from the largest-weighted (i.e., "best") particle according to

$$[\widehat{x}_k\ \widehat{y}_k]^T = [x_k^{i^*}\ y_k^{i^*}]^T, \quad i^* = \arg\max_i w_k^i \tag{7}$$

prior to resampling. Iterative application of (6) or (7) is the standard method for tracking video targets.

Tracking failures can occur for several reasons. One is that, in all but the simplest cases, the appearance model must be dynamically updated. Over time, it can become corrupted by clutter leakage and drift ultimately leading to track loss. The question of how best to perform robust appearance model updates has received considerable attention recently [5]–[7]. Another problem is that "bad" particles can sometimes receive large weights. This happens frequently in cases where there is strongly structured clutter together with significant target appearance changes. The likelihood $p(\mathbf{z}_k|\mathbf{x}_k^i)$ is increased for bad particles that have a poor state hypothesis but partially match the clutter, while $p(\mathbf{z}_k|\mathbf{x}_k^i)$ is decreased for good particles that have a good state hypothesis but only partially match the target due to an ineffective appearance model update strategy. The combined effect can be substantial. Furthermore, it is exacerbated by the fact that the decreases in $p(\mathbf{z}_k|\mathbf{x}_k^i)$ for the good particles that should be *heavily* weighted tends to increase the proportionality constant in (5), further amplifying the weights of the bad particles.

When a bad particle receives a large weight, it is turned into *many* bad particles by the SIR resampling operation. Since the total number of particles is fixed, this reduces the number of particles that are available in the next time step for searching the "good" part of the state space where the true target state lies. This generally has a deleterious effect on the tracking performance. Moreover, the presence of many particles in the "bad" parts of the state space implies that increasing numbers of bad particles may partially match the clutter and receive large weights in subsequent time steps due to the stochastic nature of $\mathbf{v}_k$ in (2). The tracker often fails when this occurs.

The importance of motion information in video tracking has been well recognized [2], [3], [5], [8]. However, in practical video tracking applications the sensor produces one video frame at each time step and there is no way to obtain direct velocity measurements without incorporating additional sensors. Thus, even if the motion model is good, direct velocity measurements are not available in $\mathbf{z}_k$ and according to (5) there is no way of using the weight calculation to explicitly penalize particles with poor velocity hypotheses. Indeed, the likelihood calculation is often implemented in a way that considers the appearance variables only and omits the velocity information altogether. This has two unfortunate consequences: 1) a particle with a state hypothesis $\mathbf{x}_k^i$ that matches the true target well will receive the same weight as a particle with an identical appearance hypothesis but a poor velocity hypothesis; and 2) even in the presence of ego motion and moving clutter, it is unlikely for the clutter to match the target in both appearance and velocity. Nevertheless, a particle that partially matches the clutter in appearance may receive a large weight even though this should be preventable based on velocity information.

Of course, one can save the previous frame and difference the positions of best matching blocks between it and the current frame to produce estimated velocity measurements. However, a naïve implementation of this approach approximates the derivative with a zeroth-order Taylor series which is notoriously poor and, more importantly, can suffer from imprecision and uncertainty in the block matching algorithm.

In this paper, we address these problems by considering the question of how to prevent particles with a poor velocity hypothesis from receiving large weights. In Section II, we introduce a new state vector augmentation method to incorporate indirect velocity information into the likelihood function. Preliminary experimental evaluation of the new method is presented in Section III, while discussion and conclusions are reserved for Section IV.

II. Incorporating Velocity Information into the Likelihood Function

In this section, we introduce our proposed method in the context of template tracking [5], [8]–[11]. The reason is that we are most interested in tracking targets in monochromatic infrared video signals that are characterized by low contrast, poor SNR, and strong clutter. Template tracking often works better than other approaches for such signals. However, the proposed method extends easily to appearance models based on, e.g., histograms, HOG, SIFT, LBP or other features.

Suppose that the sensor delivers a sequence of video frames $\mathbf{z}_k$ and let $\mathbf{T}$ be a globally defined appearance model (template) for the target. We begin by considering a very standard conventional SIR formulation where (2) takes the form shown in (8) below. In (8), Δ is the frame time and $v_k^{\cdot}$ are mutually uncorrelated i.i.d. noises that model the temporal second derivatives of position, magnification, and rotation. Note that (8) *defines* the matrix $\mathbf{A}$ and the vector $\mathbf{v}_k$. The function $\mathbf{h}$ in (3) creates a frame of zeros, inserts the target model $\mathbf{T}$ at position $[x_k \;\; y_k]^T$ with magnification γ_k and rotation θ_k, and adds measurement noise $\mathbf{n}_k$ (there is no explicit background model).

For a particle $\mathbf{x}_k^i$, the hypothesized target appearance is $\mathbf{z}_k^i = \mathbf{h}(\mathbf{x}_k^i, \mathbf{0})$. Let $\mathbf{\Omega}_k^i$ be the spatial support of the magnified and rotated template in $\mathbf{z}_k^i$. The SIR filter assigns the particle weight (5) using likelihood

$$p(\mathbf{z}_k|\mathbf{x}_k^i) = e^{-\kappa(1-\rho_k^i)}, \tag{9}$$

where κ is a gain that must be tuned and ρ_k^i is normalized cross correlation given by

$$\rho_k^i = \frac{\sum_{\mathbf{\Omega}_k^i} \mathbf{z}_k \mathbf{z}_k^i}{\sqrt{\sum_{\mathbf{\Omega}_k^i} \mathbf{z}_k^2 \sum_{\mathbf{\Omega}_k^i} (\mathbf{z}_k^i)^2}}. \tag{10}$$

Eqs. (3)-(5), (8)-(10), and (6) or (7) define the conventional SIR track filter that will be used as a baseline for comparison with the proposed method.

Let $\mathbf{x}_{k-1}^j$ be the particle from which $\mathbf{x}_k^i$ was resampled in time step $k-1$ and define $\mathbf{x}_{k-1}^{(i)} \equiv \mathbf{x}_{k-1}^j$. The likelihood function (9) makes no explicit consideration of velocity information. Furthermore, no direct measurements of velocity are available. However, if particle $\mathbf{x}_k^i$ has a good appearance hypothesis then ρ_k^i should be large. If the velocity hypothesis is also good, then ρ_{k-1}^j should also have been large. We want the particle $\mathbf{x}_k^i$ to receive a large weight only if both of these conditions are met. This suggests a way to *indirectly* incorporate velocity information into the likelihood function $p(\mathbf{z}_k|\mathbf{x}_k^i)$ if the previous measurement $\mathbf{z}_{k-1}$ can be retained in $\mathbf{z}_k$ and the previous hypothesis $\mathbf{x}_{k-1}^{(i)}$ can be retained in $\mathbf{x}_k^i$. State vector augmentation provides a means to do this. We denote quantities in the augmented system model with a tilde. With respect to (1), the augmented state vector is given by $\widetilde{\mathbf{x}}_k = [\mathbf{x}_k \;\; \mathbf{x}_{k-1}]^T$. The augmented state update model is then

$$\widetilde{\mathbf{x}}_{k+1} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \widetilde{\mathbf{x}}_k + \begin{bmatrix} \mathbf{v}_k \\ \mathbf{0} \end{bmatrix}, \tag{11}$$

where $\mathbf{I}$ is the 8×8 identity matrix. Now let $\breve{\mathbf{z}}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{n}_{k+1})$ be a second realization of $\mathbf{z}_k$ that is generated at time step $k+1$ instead of time step k. The difference between this second realization and the original measurement $\mathbf{z}_k$ is $\breve{\mathbf{z}}_k - \mathbf{z}_k = \mathbf{n}_{k+1} - \mathbf{n}_k$. We define the augmented measurement $\widetilde{\mathbf{z}}_k = [\mathbf{z}_k \;\; \breve{\mathbf{z}}_{k-1}]^T$. The augmented

$$\mathbf{x}_{k+1} = \begin{bmatrix} x_{k+1} \\ \dot{x}_{k+1} \\ y_{k+1} \\ \dot{y}_{k+1} \\ \gamma_{k+1} \\ \dot{\gamma}_{k+1} \\ \theta_{k+1} \\ \dot{\theta}_{k+1} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} & & & \mathbf{0} & & & \\ & \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} & & & & & \\ & & \mathbf{0} & & \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} & & \\ & & & & & \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} & \end{bmatrix} \begin{bmatrix} x_k \\ \dot{x}_k \\ y_k \\ \dot{y}_k \\ \gamma_k \\ \dot{\gamma}_k \\ \theta_k \\ \dot{\theta}_k \end{bmatrix} + \begin{bmatrix} 0 \\ v_k^x \\ 0 \\ v_k^y \\ 0 \\ v_k^\gamma \\ 0 \\ v_k^\theta \end{bmatrix} \equiv \mathbf{A}\mathbf{x}_k + \mathbf{v}_k \tag{8}$$

measurement model is then

$$\widetilde{\mathbf{z}}_k = \begin{bmatrix} \mathbf{z}_k \\ \breve{\mathbf{z}}_{k-1} \end{bmatrix} = \widetilde{\mathbf{h}}(\widetilde{\mathbf{x}}_k, \mathbf{n}_k) \equiv \begin{bmatrix} \mathbf{h}(\mathbf{x}_k, \mathbf{n}_k) \\ \mathbf{h}(\mathbf{x}_{k-1}, \mathbf{n}_k) \end{bmatrix}. \quad (12)$$

For an augmented particle $\widetilde{\mathbf{x}}_k^i = [\mathbf{x}_k^i \ \mathbf{x}_{k-1}^{(i)}]^T$ with measurement hypothesis $\widetilde{\mathbf{z}}_k^i = [\mathbf{z}_k^i \ \breve{\mathbf{z}}_{k-1}^i]^T$, let $\breve{\mathbf{\Omega}}_{k-1}^i$ be the spatial support of the magnified and rotated template in $\breve{\mathbf{z}}_k^i$. The likelihood may now be defined according to

$$p(\widetilde{\mathbf{z}}_k | \widetilde{\mathbf{x}}_k^i) = e^{-\kappa(1-\widetilde{\rho}_k^i)}, \quad (13)$$

where

$$\widetilde{\rho}_k^i = \frac{\sum_{\mathbf{\Omega}_k^i} \mathbf{z}_k \mathbf{z}_k^i + \sum_{\breve{\mathbf{\Omega}}_{k-1}^i} \mathbf{z}_{k-1} \breve{\mathbf{z}}_{k-1}^i}{\sqrt{\sum_{\mathbf{\Omega}_k^i} \mathbf{z}_k^2 + \sum_{\breve{\mathbf{\Omega}}_{k-1}^i} \mathbf{z}_{k-1}^2} \sqrt{\sum_{\mathbf{\Omega}_k^i} (\mathbf{z}_k^i)^2 + \sum_{\breve{\mathbf{\Omega}}_{k-1}^i} (\breve{\mathbf{z}}_{k-1}^i)^2}} \quad (14)$$

is the normalized cross correlation between the hypothesis of augmented particle $\widetilde{\mathbf{x}}_k^i$ and two consecutive video frames acquired from the camera. Intuitively, $\breve{\mathbf{z}}_{k-1}$ is nothing more than a *theoretical* construct that provides us with a rigorous way to consider a realization of $\mathbf{z}_{k-1}$ in the likelihood $p(\widetilde{\mathbf{z}}_k | \widetilde{\mathbf{x}}_k)$. Note that $\breve{\mathbf{z}}_{k-1}$ does not appear in (14); rather, it is the actual saved video frame $\mathbf{z}_{k-1}$ that is correlated with the lagged particle hypothesis $\breve{\mathbf{z}}_{k-1}^i$. Eqs. (11)-(14), (4), (5), and (6) or (7) define our new method for incorporating indirect velocity information into the SIR filter likelihood function.

III. Experiments

We compared the proposed method against the standard SIR filter on two synthetic video sequences and two of the longwave IR sequences described in [12]. Appearance model updates were not performed in order to isolate performance differences between the SIR filters with standard likelihood (9) and with velocity information incorporated via state augmentation and likelihood (13). The synthetic sequences were simulated by inserting the target of Fig. 1(a) into benign and complex backgrounds as shown in Figs. 1(c) and (b) with trajectory determined by (8). The horizontal and vertical velocity drift noises were Gaussian with variances 0.63 and 0.75, while the magnification and rotation drift noises were uniform with variances 3.6×10^{-5} and 6.4×10^{-3}. For the IR sequences shown in Figs. 1(g)-(o), ground truth was compiled manually. The noise variances were estimated by approximating derivatives with finite differences on the ground truth data – which is suboptimal as already noted but was chosen in the interest of expediency. The initial particle sets were distributed normally (position) and uniformly (magnification/rotation) with means and variances given by ground truth from the first frame.

The tracking results are given in Fig. 1 and Table I. As we expected, incorporating velocity information had negligible impact against the benign background of Figs. 1(c), (d) but provided a substantial performance gain against the complex background of Figs. 1(e), (f) where the standard SIR filter lost the target and locked onto clutter. Against IR Sequence 1, the proposed method again provided a significant advantage when the standard SIR filter became distracted by clutter as shown in Figs. 1(i), (j). As in the benign synthetic sequence, both filters performed comparably against the second IR sequence shown in Figs. 1(l)-(o).

IV. Conclusions

We introduced a new state vector augmentation method to incorporate indirect velocity information into the likelihood function of the SIR filter for video target tracking in clutter. Our main objective was to prevent particles with a poor velocity hypothesis from receiving large weights. Our preliminary experimental results suggest that the new method is effective, preventing the track filter from locking onto structured clutter in cases where the standard SIR filter fails and delivering equivalent tracking performance in cases where the standard SIR filter succeeds.

Our method is distinct from the ones in [5] and [13] which use velocity estimates computed from the measurements to adaptively change the state update and appearance models, but do not use them explicitly in the likelihood. It is also distinct from higher-order particle filters that predict the current state from multiple previous states like the one in [3] where the likelihood depends on multiple past states but not on past measurements and the one in [2] which uses agent based crowd models. Our likelihood function (13) is similar to the one in [14] in the sense that both depend explicitly on the current and past state and the current and past observation; however, the method in [14] obtains explicit velocity estimates by block/patch matching. An important component of our future work involves comprehensive performance evaluation of our proposed method relative to all of these very interesting related techniques.

References

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online non-linear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[2] W. Liu, A. B. Chan, R. W. H. Lau, and D. Manocha, "Leveraging long-term predictions and online learning in agent-based multiple person tracking," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 25, no. 3, pp. 399–410, Mar. 2015.

[3] P. Pan and D. Schonfeld, "Visual tracking using high-order particle filtering," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 51–54, Jan. 2011.

[4] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. part I: Dynamic models," *IEEE Trans. Aerospace, Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[5] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1491–1506, Nov. 2004.

[6] Q. Wang, F. Chen, W. Xu, and M. H. Yang, "Object tracking via partial least squares analysis," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4454–4465, Oct. 2012.

[7] S. Salti, A. Cavallaro, and L. D. Stefano, "Adaptive appearance modeling for video tracking: Survey and evaluation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4334–4348, Oct. 2012.

[8] J. Sullivan and J. Rittscher, "Guiding random particles by deterministic search," in *Proc. IEEE Int'l. Conf. Comput. Vis.*, vol. 1, Vancouver, BC, Canada, Jul. 7-14, 2001, pp. 323–330.

Table I
PERFORMANCE COMPARISON BETWEEN PROPOSED METHOD AND STANDARD SIR FILTER

Case	Num Frames	N_s, Num Particles	Gain κ (9),(13)	Num Runs	Mean Absolute Tracking Error (pixels)			
					Standard SIR		Proposed SIR	
					E[·] (6)	MAP (7)	E[·] (6)	MAP (7)
Synthetic/Benign Background	150	700	10	10	**1.7505**	2.0090	1.9985	2.4028
Synthetic/Complex Background	150	700	10	20	44.2557	43.7655	**2.4348**	2.5860
Real IR Sequence 1	100	700	75	7	11.7222	12.1286	**4.0832**	4.2776
Real IR Sequence 2	80	700	75	5	**2.4934**	2.5651	2.5130	2.6138
Real IR Sequence 2	80	700	100	5	2.7948	2.8300	**2.5979**	2.6337

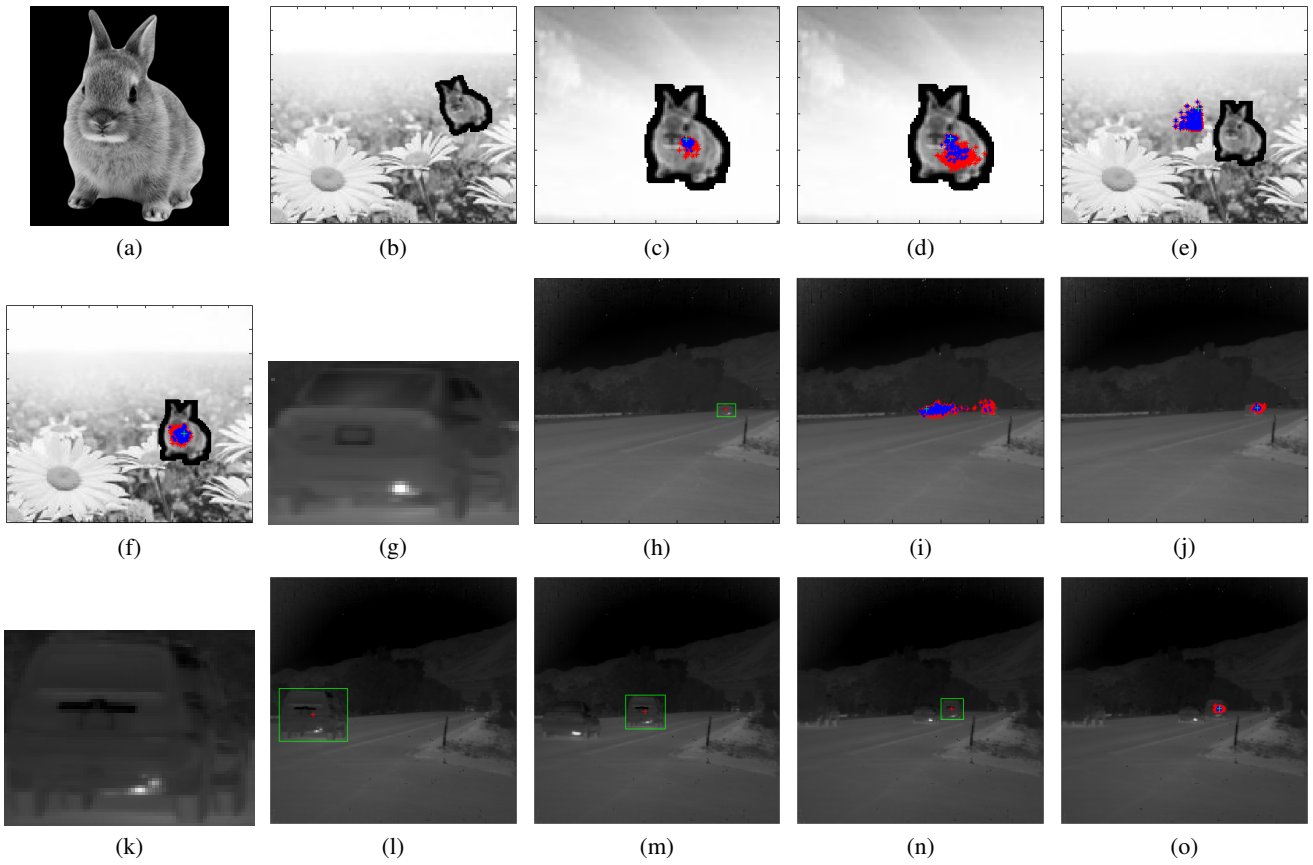

Figure 1. Examples. Where particles are shown, red/blue are the distribution before/after resampling. (a) Target for synthetic sequences. (b) Target inserted in complex background with magnification and rotation. (c),(d) Standard/proposed SIR result in benign background. (e),(f) Standard/proposed SIR result in complex background; standard SIR fails. (g) Target for Real IR Sequence 1. (h),(i),(j) Ground truth, standard SIR result, and proposed SIR result for frame 87. Standard SIR fails. (k) Target for Real IR Sequence 2. (l)-(n) ground truth for frames 2, 32, and 80. (o) Proposed SIR result for frame 81.

[9] M. G. S. Bruno, "Bayesian methods for multispectral target tracking in image sequences," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1848–1861, Jul. 2004.

[10] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Pattern Anal., Machine Intell.*, vol. 20, no. 10, pp. 1025–1039, Oct. 1998.

[11] J. Kwon, H. S. Lee, F. C. Park, and K. M. Lee, "A geometric particle filter for template-based visual tracking," *IEEE Trans. Pattern Anal., Machine Intell.*, vol. 36, no. 4, pp. 625–643, Apr. 2014.

[12] C. T. Nguyen, J. P. Havlicek, G. Fan, J. T. Caulfield, and M. S. Pattichis, "Robust dual-band MWIR/LWIR infrared target tracking," in *Proc. 48th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2-5, 2014, pp. 78–83.

[13] Y. Huang and J. Llach, "Tracking the small object through clutter with adaptive particle filter," in *Proc. Int'l. Conf. Audio, Language, Image Process.*, Jul. 7-9, 2008, pp. 357–362.

[14] J. M. Odobez, D. Gatica-Perez, and S. O. Ba, "Embedding motion in model-based stochastic tracking," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3515–3531, Nov. 2006.