2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Fusion Tracking in Color and Infrared Images Using Sequential Belief Propagation

Huaping Liu, Fuchun Sun

Abstract— **In this paper, we propose an approach to fuse the color and infrared images for visual tracking. The contribution of this paper is twofold: First, we use the covariance feature to construct the likelihood function under the framework of particle filter. This likelihood captures the spatial and statistical properties as well as their correlation within representation of covariance. Secondly, different from the existing fusion approaches, our approach automatically realizes the fusion by sequential belief propagation, which uses message passing scheme to exchange information between color and infrared image. The performance of the proposed approach is evaluated using real visual tracking examples.**

I. Introduction

Multiple sensors are able to gather more information about the reality especially in varying environmental conditions[25]. As indicated ed by [20], the rapid developments of sensor technology, microelectronics, and communications have led to a great need for image fusion techniques that can effectively combine multi-sensor images into an enhanced single view of a scene with extended information content. The use of color greatly expands the amount of information that can be conveyed in a single image and hence presents a natural approach to the representation of multi-modal data.

For fusion tracking, the choice of visual and infrared imagery is significant, as each provides disparate, yet complementary information about a scene. Infrared cameras detect relative differences in the amount of thermal energy emitted from objects in the scene. These sensors are therefore independent of illumination, making them more effective than color cameras under poor lighting conditions. Color sensors on the other hand, are oblivious to temperature differences in the scene, and are typically more effective than thermal cameras when objects are at “thermal crossover”, provided that the scene is well illuminated and the objects have color signatures different from the background[7].

Ref.[3] presented a moving object detection and tracking system that robustly fuses infrared and visible video within a level set framework. The long-term trajectories for object clusters are estimated using Kalman filtering and watershed segmentation. Kalman filtering can obtain optimal solution in the case of linear dynamics and Gaussian noise. Unfortunately, very few practical visual tracking problems belong to this case. Using mean-shift approach, [5] proposed a framework that can efficiently combine features for robust tracking based on fusing the outputs of multiple spatiogram trackers. However, mean-shift is a local deterministic search strategy, which is easy to be trapped into local minimal, and difficult to recover from tracking failure.

The particle filter, also known as sequential Monte Carlo[8], or Condensation[13], is the most popular nonlinear and non-Gaussian filtering approach. It recursively constructs the posterior probability distribution function of the state space using Monte Carlo integration. Currently, the particle filter has been extensively used in the field of location[10] and SLAM[11][1] for robots. One important advantage of the particle filtering framework is that it allows the information from different measurement sources to be fused in a principled manner[17]. During the past decade, particle filter is very popular in fusion tracking. Most of the existing works which used multiple observation information combined different observation models (or cues) in one single images[2]. Usually the combination can be realized by using a simple weighted sum form. The weighting coefficient can be determined by experience[26], or by online computation[16] and therefore forms a class of adaptive weighting approach. The approach in [16] is based on an extension of the covariance-based uncertainty measure.

Recently, using particle filter to fuse the color and infrared images attracted many attentions. [4] evaluated the appearance tracking performance of multiple fusion schemes that combine information from color and thermal infrared images for the tracking of surveillance. In [4], some common fusion algorithms are summarized and compared, including simple and weighted averaging, similarity score product, min and max score fusion and dynamic weighting approaches. [6] investigates the impact of pixel-level fusion of videos from visible and infrared surveillance cameras on object tracking performance, as compared to tracking in single modality videos. Tracking has been accomplished by means of a particle filter which fuses a color cue and the structural similarity measure.

In this paper, we propose a new approach to fuse the color and infrared images for visual tracking. The contributions of this paper is twofold: First, we will use the covariance feature to construct the likelihood function of particle filter. The covariance feature was first used for tracking in [19], which adopted the extensive search strategy. The use of this feature in particle filter has never been reported. Secondly, different from the existing fusion approaches[4], our approach automatically realizes the fusion by sequential belief propagation, which uses message passing scheme to exchange information between color and infrared images.

Huaping Liu and Fuchun Sun are both with Department of Computer Science and Technology, Tsinghua University, and State Key Laboratory of Intelligent Technology and Systems, Beijing, P.R.China.

978-1-4244-1647-9/08/$25.00 ©2008 IEEE.

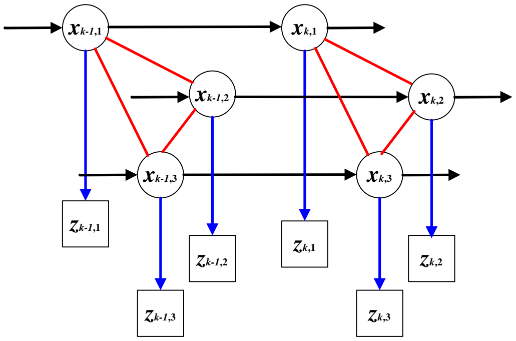

Fig. 1. Dynamic graphical model

II. Review of Sequential Belief Propagation

Given L observation images, where L is the number of the information sources, then we can denote the object state in each observed image as $\mathbf{x}_{k,i}$, where $i \in \{1, 2, \cdots, L\}$. All of the state variables can be integrated as a new variable $\mathbf{X}_k = \{\mathbf{x}_{k,1}, \mathbf{x}_{k,2}, \cdots, \mathbf{x}_{k,L}\}$. The image observation associated with the object state $\mathbf{x}_{k,i}$ in the same image is denoted by $\mathbf{z}_{k,i}$ and all of them can be integrated as $\mathbf{Z}_k = \{\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, \cdots, \mathbf{z}_{k,L}\}$.

Fig.1 gives a representative dynamic graphical model. Assume there are undirected links which describe the mutual influence of multiple information sources and it is associated with a potential function $\psi_{ij}(\mathbf{x}_{k,i}, \mathbf{x}_{k,j})$. Each directed link from $\mathbf{x}_{k,i}$ to $\mathbf{z}_{k,i}$ is associated with an image likelihood function $p(\mathbf{z}_{k,i}|\mathbf{x}_{k,i})$. In addition, the directed link from $\mathbf{x}_{k-1,i}$ to $\mathbf{x}_{k,i}$ represents the prior dynamics and is associated with a dynamics model $p(\mathbf{x}_{k,i}|\mathbf{x}_{k-1,i})$. According to Bayes' rule, the recursive inference of the posterior distribution of the state $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$, where $\mathbf{Z}_{1:k} = [\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_k]$, is formulated as

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k-1}) = \int_{\mathbf{X}_{k-1}} p(\mathbf{X}_k|\mathbf{X}_{k-1}) p(\mathbf{X}_{k-1}|\mathbf{Z}_{1:k-1}) d\mathbf{X}_{k-1} \quad (1)$$

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) \propto p(\mathbf{Z}_k|\mathbf{X}_k) p(\mathbf{X}_k|\mathbf{Z}_{1:k-1}) \quad (2)$$

The inference of the joint multi-source state is difficult. To tackle it, the sequential belief propagation approach can be adopted. The semi-parametric sequential belief propagation was first proposed by [21] and [14]. After then Ref. [12] proposed the non-parametric sequential belief propagation for multiple scale tracking. Recently, [9] and [15] used this approach for multi view tracking and head-face fusion tracking, respectively. In this section, we will briefly review this approach, which is based on a basic assumption that the motions in different images are independent, i.e.,

$$p(\mathbf{X}_k|\mathbf{X}_{k-1}) = \prod_{i=1}^{L} p(\mathbf{x}_{k,i}|\mathbf{x}_{k-1,i}) \quad (3)$$

The intrinsic idea of the multi-source tracking algorithm is to calculate the inference of states through a message passing process. The local message passed from source i to source j in the graphical model in Fig.1 is

$$\begin{aligned} m_{ij}(\mathbf{x}_{k,j}) = \int_{\mathbf{x}_{k,i}} [&p(\mathbf{z}_{k,i}|\mathbf{x}_{k,i})\psi_{i,j}(\mathbf{x}_{k,i}, \mathbf{x}_{k,j}) \\ &\int_{\mathbf{x}_{k-1,i}} p(\mathbf{x}_{k,i}|\mathbf{x}_{k-1,i}) p(\mathbf{x}_{k-1,i}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,i} \\ &\prod_{l\in\mathcal{N}(\mathbf{x}_{k,i})\backslash j} m_{li}(\mathbf{x}_{k,i})] d\mathbf{x}_{k,i} \end{aligned} \quad (4)$$

where $\mathcal{N}(\mathbf{x}_{k,i})\backslash j$ denotes all state variables with a link to $\mathbf{x}_{k,i}$, except $\mathbf{x}_{k,j}$. The messages are passed iteratively until convergence, and the filtering distribution is given by

$$\begin{aligned} p(\mathbf{x}_{k,i}|\mathbf{Z}_{1:k}) \propto\ & p(\mathbf{z}_{k,i}|\mathbf{x}_{k,i}) \prod_{l\in\mathcal{N}(\mathbf{x}_{k,i})} m_{li}(\mathbf{x}_{k,i}) \\ & \int_{\mathbf{x}_{k-1,i}} p(\mathbf{x}_{k,i}|\mathbf{x}_{k-1,i}) p(\mathbf{x}_{k-1,i}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,i} \end{aligned} \quad (5)$$

Since the closed-form solutions to the two distribution are difficult to obtain, a Monte Carlo version of sequential belief propagation is developed. The filtering distribution is represented by weighted samples, i.e., $p(\mathbf{x}_{k,i}|\mathbf{Z}_{1:k}) \sim \{\mathbf{x}_{k,i}^{(n)}, \pi_{k,i}^{(n)}\}_{n=1}^{N}$, and each message at time instant k can be approximated as $m_{ji}(\mathbf{x}_{k,i}) \sim \{\mathbf{x}_{k,i}^{(n)}, \omega_{k,i}^{(n)}\}_{n=1}^{N}$, where N is the number of the particles. In the following section, we will introduce the application of the general sequential belief propagation algorithm in the fusion of color and infrared images.

III. Sequential Belief Propagation for Color and Infrared Images Fusion

Given two information sources: color and infrared images, we can represent the object state in the two images as $\mathbf{x}_{k,C} = [x_{k,C}\ \ y_{k,C}\ \ s_{k,C}]^T$, $\mathbf{x}_{k,I} = [x_{k,I}\ \ y_{k,I}\ \ s_{k,I}]^T$, where $\{x_{k,C}, y_{k,C}\}$ and $\{x_{k,I}, y_{k,I}\}$ represent the center coordinates of the object in color and infrared images, respectively. $s_{k,C}$ and $s_{k,I}$ represent the corresponding scales. If there is no registration error, $\mathbf{x}_{k,C}$ and $\mathbf{x}_{k,I}$ should be equal since they represent the state variable of the same object; Otherwise they are different.

Since there has only two state variable nodes in the graphic model, (4) and (5) can be simplified as

$$\begin{aligned} m_{IC}(\mathbf{x}_{k,C}) = \int_{\mathbf{x}_{k,I}} [&p(\mathbf{z}_{k,I}|\mathbf{x}_{k,I})\psi_{I,C}(\mathbf{x}_{k,I}, \mathbf{x}_{k,C}) \\ &\int_{\mathbf{x}_{k-1,I}} p(\mathbf{x}_{k,I}|\mathbf{x}_{k-1,I}) p(\mathbf{x}_{k-1,I}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,I}] d\mathbf{x}_{k,I} \\ m_{CI}(\mathbf{x}_{k,I}) = \int_{\mathbf{x}_{k,C}} [&p(\mathbf{z}_{k,C}|\mathbf{x}_{k,C})\psi_{C,I}(\mathbf{x}_{k,C}, \mathbf{x}_{k,I}) \\ &\int_{\mathbf{x}_{k-1,C}} p(\mathbf{x}_{k,C}|\mathbf{x}_{k-1,C}) p(\mathbf{x}_{k-1,C}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,C}] d\mathbf{x}_{k,C} \end{aligned} \quad (6)$$

and

$$\begin{aligned} p(\mathbf{x}_{k,C}|\mathbf{Z}_{1:k}) \propto\ & p(\mathbf{z}_{k,C}|\mathbf{x}_{k,C}) m_{IC}(\mathbf{x}_{k,C}) \\ & \int_{\mathbf{x}_{k-1,C}} p(\mathbf{x}_{k,C}|\mathbf{x}_{k-1,C}) p(\mathbf{x}_{k-1,C}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,C} \\ p(\mathbf{x}_{k,I}|\mathbf{Z}_{1:k}) \propto\ & p(\mathbf{z}_{k,I}|\mathbf{x}_{k,I}) m_{CI}(\mathbf{x}_{k,I}) \\ & \int_{\mathbf{x}_{k-1,I}} p(\mathbf{x}_{k,I}|\mathbf{x}_{k-1,I}) p(\mathbf{x}_{k-1,I}|\mathbf{Z}_{1:k-1}) d\mathbf{x}_{k-1,I} \end{aligned} \quad (7)$$

In the fusion of color and infrared images, the potential function $\psi_{C,I}(\mathbf{x}_{k,C}, \mathbf{x}_{k,I})$ and $\psi_{I,C}(\mathbf{x}_{k,I}, \mathbf{x}_{k,C})$ model the registration error. In our case, we use a Gaussian distribution to represent it, i.e.,

$$\begin{aligned}\psi_{C,I}(\mathbf{x}_{k,C}, \mathbf{x}_{k,I}) &= \psi_{I,C}(\mathbf{x}_{k,I}, \mathbf{x}_{k,C}) \\ &\propto \exp(-(\mathbf{x}_{k,C} - \mathbf{x}_{k,I})^T \Lambda (\mathbf{x}_{k,C} - \mathbf{x}_{k,I}))\end{aligned} \tag{8}$$

where Λ is the variance matrix, which can be determined by the registration process. The whole sequential belief propagation algorithm, which is similar to that of [9][12][15], is summarized in **Algorithm 1**.

IV. Observation likelihood function

A major concern is the lack of a competent similarity criterion that captures both statistical and spatial properties, i.e., most approaches either depend only on the color distributions or structural models. Many different representations, from aggregated statistics to appearance models, have been used for tracking objects. Color histograms are popular representations of nonparametric density, but they disregard the spatial arrangement of the feature values. Moreover, they do not scale to higher dimensions due to exponential size and sparsity. Appearance models map the image features onto a fixed size window. Since the dimensionality is a polynomial in the number of features and the window size, only a relatively small number of features can be used. Appearance models are highly sensitive to the pose, scale and shape variations. To overcome the shortcomings of the existing approaches, Ref.[18] proposed a covariance matrix representation to describe the object windows. Ref.[22] used this approach for detection and classification. In [19], this approach is extended to tracking domain. Recently, [23] combined the Logitboost and the covariance feature to detect humans in still images. However, till now, there exists no work to use the covariance feature for constructing likelihood function for particle filter.

For a given rectangular region which includes M_r rows and M_c columns, let $\mathbf{f}_{ij} (i = 1, \cdots, M_r, j = 1, \cdots, M_c)$ be the d-dimensional feature vectors inside this region for each pixel. The feature vector $\mathbf{f}_{ij}$ can be constructed using two types of mappings: spatial attributes that are obtained from pixel coordinate values, and appearance attributes, i.e., color, gradient, etc. These features may be associated directly to the pixel coordinates.

The covariance feature of a region can be calculated as

$$\mathbf{C} = \frac{1}{M_r M_c - 1} \sum_{i=1}^{M_r} \sum_{j=1}^{M_c} (\mathbf{f}_{ij} - \mu)(\mathbf{f}_{ij} - \mu)^T$$

where μ is the vector of the means of the corresponding features for the points within this region. The covariance matrix is a symmetric matrix where its diagonal entries represent the variance of each feature and the non-diagonal entries represent their respective correlations.

Supposing no features in the feature vector would be exactly identical, which states the covariance matrices are positive definite, it is possible apply the distance measure. The distance metric uses the sum of the squared logarithms of the generalized eigenvalues to compute the dissimilarity

Algorithm 1 `Fusion tracking algorithm`

```
Given: ${\mathbf{x}_{k,C}^{(n)}, \pi_{k,C}^{(n)}\}_{n=1}^N, \{\mathbf{x}_{k,I}^{(n)}, \pi_{k,I}^{(n)}\}_{n=1}^N$
Output: $\{\mathbf{x}_{k+1,C}^{(n)}, \pi_{k+1,C}^{(n)}\}_{n=1}^N, \{\mathbf{x}_{k+1,I}^{(n)}, \pi_{k+1,I}^{(n)}\}_{n=1}^N$

Re-sampling: For $n = 1, 2, \cdots, N$, resampling $\{\mathbf{x}_{k,C}^{(n)}, \pi_{k,C}^{(n)}\}_{n=1}^N, \{\mathbf{x}_{k,I}^{(n)}, \pi_{k,I}^{(n)}\}_{n=1}^N$ to get $\{\hat{\mathbf{x}}_{k,C}^{(n)}, 1/N\}_{n=1}^N, \{\hat{\mathbf{x}}_{k,I}^{(n)}, 1/N\}_{n=1}^N$

Prediction: For $n = 1, 2, \cdots, N$, draw predicted particles from the prior dynamics
    $\mathbf{x}_{k+1,C}^{(n)} \sim p(\mathbf{x}_{k+1,C} | \mathbf{x}_{k,C} = \hat{\mathbf{x}}_{k,C}^{(n)})$,
    $\mathbf{x}_{k+1,I}^{(n)} \sim p(\mathbf{x}_{k+1,I} | \mathbf{x}_{k,I} = \hat{\mathbf{x}}_{k,I}^{(n)})$,
and initialize
    $\pi_{k+1,C}^{(n)} = p(\mathbf{z}_{k+1,C} | \mathbf{x}_{k+1,C}^{(n)}), \pi_{k+1,I}^{(n)} = p(\mathbf{z}_{k+1,I} | \mathbf{x}_{k+1,I}^{(n)})$,
    $\omega_{k+1,C}^{(n)} = 1/N, \omega_{k+1,I}^{(n)} = 1/N.$
Update:
(U.1) Importance Sampling
    $\bar{\mathbf{x}}_{k+1,C}^{(n)} \sim p(\mathbf{x}_{k+1,C} | \mathbf{x}_{k,C} = \hat{\mathbf{x}}_{k,C}^{(n)})$
    $\bar{\mathbf{x}}_{k+1,I}^{(n)} \sim p(\mathbf{x}_{k+1,I} | \mathbf{x}_{k,I} = \hat{\mathbf{x}}_{k,I}^{(n)})$
(U.2) Message Re-weighting
    $\omega_{k+1,C}^{(n)} = G_C / (\frac{1}{N} \sum_{r=1}^N p(\bar{\mathbf{x}}_{k+1,C}^{(n)} | \hat{\mathbf{x}}_{k,C}^{(r)}))$
    $\omega_{k+1,I}^{(n)} = G_I / (\frac{1}{N} \sum_{r=1}^N p(\bar{\mathbf{x}}_{k+1,I}^{(n)} | \hat{\mathbf{x}}_{k,I}^{(r)}))$
where
    $G_C = \sum_{m=1}^N \pi_{k+1,I}^{(m)} p(\mathbf{z}_{k+1,I} | \mathbf{x}_{k+1,I}^{(m)}) \times \psi_{I,C}(\mathbf{x}_{k+1,I}^{(m)}, \bar{\mathbf{x}}_{k+1,C}^{(n)}) [\frac{1}{N} \sum_{r=1}^N p(\mathbf{x}_{k+1,I}^{(m)} | \mathbf{x}_{k,I}^{(r)})]$
    $G_I = \sum_{m=1}^N \pi_{k+1,C}^{(m)} p(\mathbf{z}_{k+1,C} | \mathbf{x}_{k+1,C}^{(m)}) \times \psi_{C,I}(\mathbf{x}_{k+1,C}^{(m)}, \bar{\mathbf{x}}_{k+1,I}^{(n)}) [\frac{1}{N} \sum_{r=1}^N p(\mathbf{x}_{k+1,C}^{(m)} | \mathbf{x}_{k,C}^{(r)})]$
(U.3) State Re-weighting
Normalize $\omega_{k+1,C}^{(n)}$, $\omega_{k+1,I}^{(n)}$ and set
    $\pi_{k+1,C}^{(n)} = p(\mathbf{z}_{k+1,C}^{(n)} | \bar{\mathbf{x}}_{k+1,C}^{(n)}) \sum_{r=1}^N p(\mathbf{x}_{k+1,C}^{(n)} | \bar{\mathbf{x}}_{k,C}^{(r)})$
    $\pi_{k+1,I}^{(n)} = p(\mathbf{z}_{k+1,I}^{(n)} | \bar{\mathbf{x}}_{k+1,I}^{(n)}) \sum_{r=1}^N p(\mathbf{x}_{k+1,I}^{(n)} | \bar{\mathbf{x}}_{k,I}^{(r)})$
and normalized them.
(U.4) Iteration:
    $\mathbf{x}_{k+1,C}^{(n)} \leftarrow \bar{\mathbf{x}}_{k+1,C}^{(n)}, \mathbf{x}_{k+1,I}^{(n)} \leftarrow \bar{\mathbf{x}}_{k+1,I}^{(n)}.$
Iterate (U.1) to (U.4) until convergence.
```

between covariance matrices as

$$\rho(\mathbf{C}, \bar{\mathbf{C}}) = \sqrt{\sum_{t=1}^{d} \ln^2 \lambda_t(\mathbf{C}, \bar{\mathbf{C}})}$$

where $\{\lambda_t(\mathbf{C}, \bar{\mathbf{C}})\}$ are the generalized eigenvalues of $\mathbf{C}$ and $\bar{\mathbf{C}}$, computed from

$$\lambda_t \mathbf{C}\mathbf{s}_t - \bar{\mathbf{C}}\mathbf{s}_t = 0, \qquad t = 1, 2, \cdots, d$$

and $\mathbf{s}_t$ are the corresponding generalized eigenvectors. The distance measure ρ satisfies the metric axioms, positivity, symmetry, triangle inequality, for positive definite symmetric matrices.

Denote the covariance feature of the regions determined by the state $\mathbf{x}_{k,C}$ and $\mathbf{x}_{k,I}$ as $\mathbf{C}(\mathbf{x}_{k,C})$ and $\mathbf{C}(\mathbf{x}_{k,I})$, respectively, we can compute the likelihood function as

$$p(\mathbf{z}_{k,C}|\mathbf{x}_{k,C}) \propto \exp(-\beta_C \cdot \rho(\mathbf{C}(\mathbf{x}_{k,C}), \bar{\mathbf{C}}_C))$$

$$p(\mathbf{z}_{k,I}|\mathbf{x}_{k,I}) \propto \exp(-\beta_I \cdot \rho(\mathbf{C}(\mathbf{x}_{k,I}), \bar{\mathbf{C}}_I))$$

where the subscript "C" and "I" indicate the color and infrared images, respectively. $\bar{\mathbf{C}}_C$ and $\bar{\mathbf{C}}_I$ are corresponding reference covariance model, which can be determined at the first frame. β_C and β_I are prescribed parameters. In our settings, we choose $\beta_C = \beta_I = 20$.

It is possible to compute covariance matrix from feature images in a very fast way using integral image representation. After constructing tensors of integral images corresponding to each feature dimension and multiplication of any two feature dimensions, the covariance matrix of any arbitrary rectangular region can be computed independent of the region size. Refer to [22] for more details.

It should be noted that though [19] used the covariance feature for tracking, their approach is based on an extensive search. In this paper, to the best of our knowledge, this is the first time for application of the covariance feature in the particle filtering framework.

V. Experimental results

The proposed approach is tested on thermal/color video sequence pairs from OTCBVS dataset collection[27]. Data consists of 8-bit greyscale bitmap thermal images, and 24-bit color bitmap images of 320 × 240 pixels.

For all of the experiments, the states of the particle filter are defined as $\mathbf{x}_k^{(C)} = [x_k^{(C)}, y_k^{(C)}, s_k^{(C)}], \mathbf{x}_k^{(I)} = [x_k^{(I)}, y_k^{(I)}, s_k^{(I)}]$, where $x_k^{(C)}, y_k^{(C)}$ and $x_k^{(I)}, y_k^{(I)}$ indicate the locations of the object in color and infrared images, respectively; $s_k^{(C)}$ and $s_k^{(I)}$ are the corresponding scales. The dynamics of the objects are assumed to be a random walking model, which can be represented as

$$\mathbf{x}_k^{(C)} = \mathbf{x}_{k-1}^{(C)} + \mathbf{v}_k^{(C)}, \qquad \mathbf{x}_k^{(I)} = \mathbf{x}_{k-1}^{(I)} + \mathbf{v}_k^{(I)},$$

where $\mathbf{v}_k^{(C)}$ and $\mathbf{v}_k^{(C)}$ are multivariate zero-mean Gaussian random variables. For each particle filter, we assign 100 samples.

We can initialize the particle filters with a detector algorithm[24] or a manfully specified image patch in the first frame. The covariance models for the object are extracted in the first frame and remain fixed for the duration of the experiment. For fair comparison, all of the particle filters for the sequence are started with same initial detection results.

In this experiment, we attempt to track a woman in dark clothing through occlusion and distraction by crowds. In order to examine how can the proposed fusion approach improve the tracking performance, we compare the tracking results of five algorithms. For notational simplicity, we call the five algorithms as "**Color**", "**Infrared**", "**Average**", "**Covariance**", and "**BP**", respectively. They are explained as follows:

"**Color**"

This approach uses color information only and the corresponding feature vectors are defined as

$$\mathbf{f}_{ij}^{(C)} = [i, j, R(i,j), G(i,j), B(i,j), |I_x^{(C)}(i,j)|, |I_y^{(C)}(i,j)|]$$

where i, j are pixel coordinates, $R(i,j), G(i,j), B(i,j)$ are corresponding R, G, B values, respectively; and $I_x^{(C)}(i,j), I_y^{(C)}(i,j)$ are the intensity derivatives of the grayscale version of the color images.

"**Infrared**"

This approach uses infrared information only and the corresponding feature vectors are defined as

$$\mathbf{f}_{ij}^{(I)} = [i, j, I^{(I)}(i,j), |I_x^{(I)}(i,j)|, |I_y^{(I)}(i,j)|]$$

where i, j are pixel coordinates, $I^{(I)}(i,j)$ is the corresponding grayscale value, and $I_x^{(I)}(i,j), I_y^{(I)}(i,j)$ are the intensity derivatives.

Obviously, the above two approaches do not use the fusion mechanism. In the following we will introduce two existing representative fusion approaches.

"**Average**"

This approach performs weighting average of the likelihood functions.

$$p(\mathbf{z}_k|\mathbf{x}_{k,C}, \mathbf{x}_{k,I}) = \alpha p(\mathbf{z}_{k,C}|\mathbf{x}_{k,C}) + (1-\alpha) p(\mathbf{z}_{k,I}|\mathbf{x}_{k,I})$$

where $p(\mathbf{z}_{k,C}|\mathbf{x}_{k,C})$ and $p(\mathbf{z}_{k,I}|\mathbf{x}_{k,I})$ are determined by the above-mentioned $\mathbf{f}_{ij}^{(C)}$ and $\mathbf{f}_{ij}^{(I)}$, respectively, and α is a weighting factor. In this paper, we select $\alpha = 0.5$. This setting means that we treat the color information and infrared information equally.

"**Covariance**"

This approach realizes fusion during constructing feature vectors. It was first proposed in [19] and showed good performance for fusion tracking of color and infrared images. In our case, the feature vector is defined as

$$\mathbf{f}_{ij} = \Big[\begin{array}{cccccc} i & j & R(i,j) & G(i,j) & B(i,j) & I_x^{(C)}(i,j) \\ & I_y^{(C)}(i,j) & I^{(I)}(i,j) & I_x^{(I)}(i,j) & I_y^{(I)}(i,j) & \end{array} \Big]$$

This vector is actually a combination of the above-mentioned $\mathbf{f}_{ij}^{(C)}$ and $\mathbf{f}_{ij}^{(I)}$. Adopting this covariance representation implies that $\mathbf{x}_{k,C} = \mathbf{x}_{k,I}$. Denote the corresponding covariance and reference covariance to be $\mathbf{C}(\mathbf{x}_{k,C}, \mathbf{x}_{k,I})$ and $\bar{\mathbf{C}}_{CI}$, respectively, then the likelihood is

$$p(\mathbf{z}_k|\mathbf{x}_{k,C}, \mathbf{x}_{k,I}) \propto \exp(-\beta \cdot \rho(\mathbf{C}(\mathbf{x}_{k,C}, \mathbf{x}_{k,I}), \bar{\mathbf{C}}_{CI}))$$

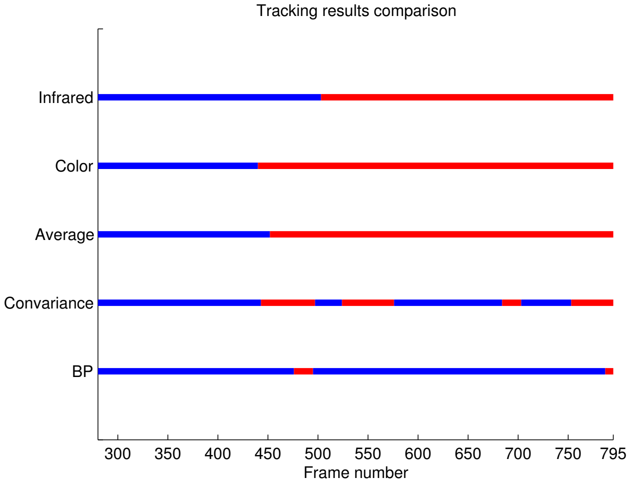

Fig. 2. Tracking performance comparisons between five algorithms. Blue dots indicate successful tracking; Red dots indicate failure tracking.

where $\beta = 20$.

"**BP**"

This is the proposed approach in this paper. The utilized likelihood function is also determined by the above-defined $\mathbf{f}_{ij}^{(C)}$ and $\mathbf{f}_{ij}^{(I)}$. The other details are omitted.

In Fig.2, we give the comparison between the performances of different algorithms, where the successful tracked frames are marked with blue and otherwise are marked with red. A more quantificational comparison will be our future work.

From Fig.2 we can see all of the algorithms failed during Frames 476-495. In fact, during this period, there is a telegraph pole which occludes the object and therefore the tracking performances are decreased. However, after the object re-appears (about Frame 495), the approaches "**Color**" and "**Average**" never recovers the accurate position of the object and locks onto the telegraph pole. The approach "**Infrared**", which uses infrared information only, can recover the tracking of the object. However, from Frame 503, the tracker will be hijacked by a passing person which is similar to the tracked object and gives the wrong tracking results. It seems only the approaches "**Covariance**" and "**BP**" can recover the performance from the influence of the telegraph pole, and can resist the attraction of the passing person. Furthermore, from this figure we can see the approach "**BP**" tracks the person almost throughout the entire sequence, despite severe occlusion and background distraction, while the approaches "**Covariance**" fails during some short periods. Finally, we notice that the approach "**BP**" also fail during Frames 787-795(the last frame). This is due to the reason that during these frames the tracked person is occluded by some trees, which produce clutter.

Figs.3-5 give some representative examples. Fig.3 is for the approach "**Infrared**". During Frames 440-467, the object walks near the telegraph pole, and the tracking performance is satisfactory, though the telegraph pole partially occludes the object. During Frames 467-504, the object re-appears and the occluding finishes, the tracker recovers the accurate tracking of the object. However, at Frame 505, the tracker is hijacked by a near passing person which is similar to the object and from then on the tracker locks onto this person; Therefore it totally fail during Frame 505 to the end(See Fig.3 for the representative frames 511 and 564).

On the other hand, the approach "**Average**", though fuses the information of color and infrared, fails at Frame 452. After that frame it locks onto the telegraph pole and never recovers. This shows that simple averaging can not provide robust fusion for this case. See Fig.4 for representative frames.

Finally, we compare the tracking results of the approaches "**Covariance**" and "**BP**". From Fig.2 we see that "**BP**" can provide more stable tracking performance than "**Covariance**". This can be shown from Fig.5. We can see the proposed approach can give better performance than the approach "**Covariance**". More results are omitted due to the page limitations.

VI. Conclusions

The contribution of this paper is twofold: First, the covariance feature is used to construct the likelihood function under the framework of particle filter. This likelihood captures the spatial and statistical properties as well as their correlation within representation of covariance. Secondly, different from the existing fusion approaches, our approach automatically realize the fusion by sequential belief propagation, which uses message passing scheme to exchange information between color and inferred image information. The performance of the proposed approach is evaluated using real visual tracking examples.

VII. Acknowledgements

This work was jointly supported by the National Science Fund for Distinguished Young Scholars (Grant No: 60625304), the National Natural Science Foundation of China (Grants No: 60504003, 60474025), and the National Key Project for Basic Research of China (Grant No: 2007CB311003).

References

[1] J.L.Blanco, J.A.Fernandez, J.Gonzalez, An entropy-based measurement of certainty in Rao-Balckwelized particle filter mapping, *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006, pp.3550-3555

[2] L. Brethes, F.Lerasle, P.Danes, Data fusion for visual tracking dedicated to human-robot interaction, in: *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2005, pp.2075-2080

[3] F. Bunyak, K. Palaniappan, S. K. Nath, G. Seetharaman, Geodesic active contour based fusion of visible and infrared video for persistent object tracking, in: *Proc. of IEEE Workshop on Applications of Computer Vision (WACV)*, 2007, pp.35-35

[4] C. Ò. Conaire, N. E. O'Connor, E. Cooke, A. F. Smeaton, Comparison of fusion methods for thermo-visual surveillance tracking, in: *Proc. of Int. Conf. on Information Fusion*, 2006, pp.1-7

[5] C. Ò. Conaire, N. E. O'Connor, A. F. Smeaton, Thermo-visual feature fusion for object tracking using multiple spatiogram trackers, *Machine Vision and Applications*, 2007

[6] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, C. N. Canagarajah, The effect of pixel-level fusion on object tracking in multi-sensor surveillance video, in: *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2007, pp.1-7

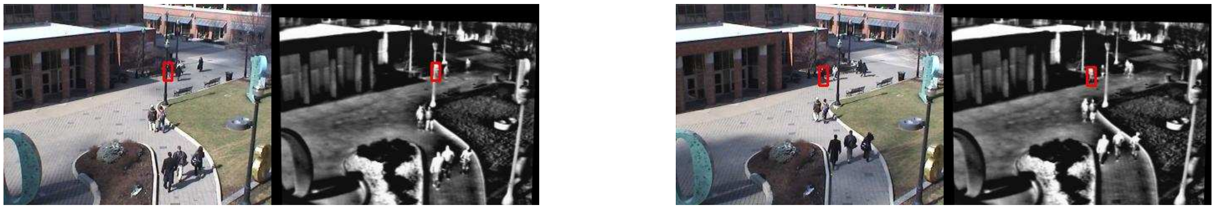

Fig. 3. The approach “**Infrared**”. Left: Frame 511; Right: Frame 564. (Note that the passing person walks to the left and the true object walks to the right. The tracker locks onto the passing person and the tracking totally fails.)

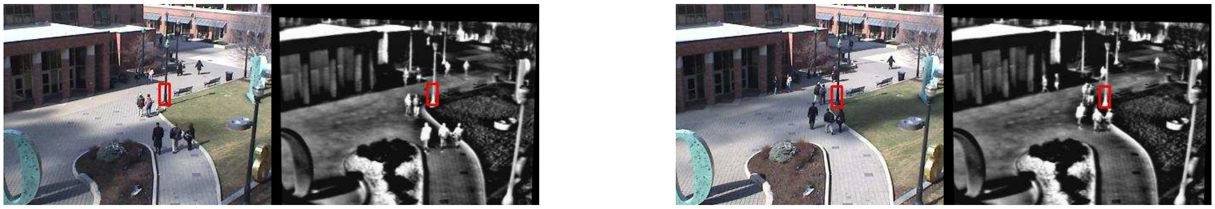

Fig. 4. The approach “**Average**”. Left: Frame 601; Right: 691. (After the object leaves the telegraph pole and re-appears, the tracking results never recover.)

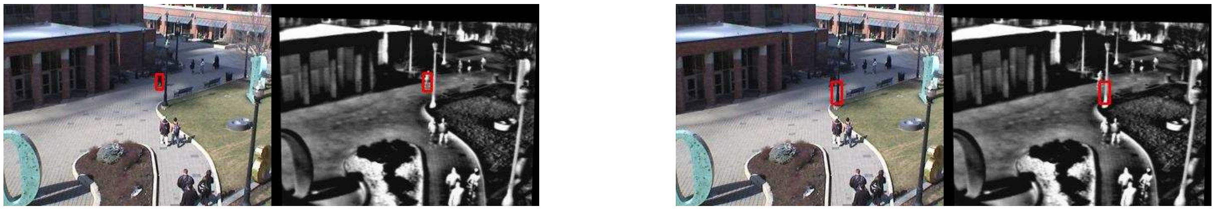

Fig. 5. Frame 446. Left: “**BP**”; Right: “**Covariance**”

[7] J. Davis, V. Sharma. Fusion-based background subtraction using contour saliency, in: *Proc. of IEEE Int. Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, San Diego, CA, June 2005, pp.1-8

[8] A. Doucet, N. de Freitas, N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001

[9] W. Du, J. Piater, Multi-view object tracking using sequential belief propogation, in: *Proc. of Asian Conf. on Computer Vision (ACCV)*, 2006, pp.684-693

[10] A. Gning, F. Abdallah, P. Bonnifait, A new estimation method for multisensor fusion by using interval analysis and particle filtering, in: *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2007, pp.3844 - 3849

[11] D. Hahnel, W. Burgard, D. Fox, S. Thrun, An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements, in: *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2003, pp.206-211

[12] G. Hua, Y. Wu, Multi-scale visual tracking by sequential belief propogation, in: *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2004, pp.826-833

[13] M. Isard, A. Blake, Condensation - conditional desity propagation for visual tracking, *Int. J. of Computer Vision*, vol.29, no.1, 1998, pp.5-28

[14] M. Isard, PAMPAS: Real-valued graphical models for computer vision, in: *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2003, pp.613-620

[15] Y. Li, H. Ai, C. Huang, S. Lao, Robust head tracking based on a multi-state particle filter, in: *Proc. of Face and Gesture Recognition (FGR)*, 2006, pp.335-340

[16] E. Maggio, F. Smeraldi, A. Cavallaro, Combining colour and orientation for adaptive particle filter-based tracking. in: *Proc. of British Machine Vision Conference*, 2005, pp.1-10

[17] P. Perez, J. Vermaak, A. Blake, Data fusion for visual tracking with particles, *Proc. of IEEE*, vol.92, no.3, 2004, pp.495-513

[18] F. Porikli, O. Tuzel, Fast construction of covariance matrices for arbitrary size image windows, in: *Proc. of Int. Conf. on Image Processing (ICIP)*, 2006, pp.1581-1584

[19] F. Porikli, O. Tuzel, P. Meer, Covariance tracking using model update based on Lie algebra, in: *Proc. of Computer Vision and Pattern Recogntion (CVPR)*, 2006, pp.728-735

[20] M. I. Smith, J. P. Heather, A review of image fusion technology in 2005, in: *Proc. of SPIE*, vol.5782, 2005, pp.29-45

[21] E. B. Sudderth, A. T. Ihler, W. T. Freeman, A. S. Willsky, Nonparametric belief propagation, in: *Proc. of Computer Vision and Pattern Recogntion (CVPR)*, 2003, pp.605-612

[22] O. Tuzel, F. Porikli, P. Meer, Region covariance: A fast descriptor for detection and classification, in: *Proc. of European Conf. on Computer Vision (ECCV)*, 2006, pp.589-600

[23] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on Riemannian manifolds, in: *Proc. of Computer Vision and Pattern Recogntion (CVPR)*, 2007, pp.1-8

[24] P.Viola, M.J.Jones, Robust real-time face detection, *Int. J. Computer Vision*, vol.52, no.2, 2004, pp.137-154

[25] L. Walchshäusl, R. Lindl, Multi-sensor classification using a boosted cascade detector, in: *Proc. of the IEEE Intelligent Vehicles Symposium*, 2007, pp.1045-1049

[26] X. Xu, B. Li, Head tracking using particle filter with intensity gradient and color histogram, *IEEE International Conference on Multimedia and Expo (ICME)*, 2005, pp.888-891

[27] http://www.cse.ohio-state.edu/OTCBVS-BENCH