Fusion by Optimal Dynamic Mixtures of Proposal Distributions

Tony X. Han*, Huazhong Ning†, Thomas S. Huang†

*ECE Department
University of Missouri
349 Engr. Bldg. W., Columbia, MO 65211
hantx@missouri.edu

†Beckman Institute and ECE Department
University of Illinois
405 N. Mathews Ave., Urbana, IL 61801
{*hning2, huang*}*@ifp.uiuc.edu*

Abstract

We propose a fusion framework to integrate multiple cues for tracking by finding a set of optimal dynamic weights for different tracking modalities. In the setup of Bayesian sequential estimation, we give an optimal criterion to find the dynamic weight for each modality: Using a linear combination of the proposal distributions from multiple cues to approach the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_t)$. The fusion problem is then formulated as an optimization problem with a non-convex objective function. We further convert the optimization problem to a constrained convex programming problem. The equations for finding the global optimal solution are given and an approximate analytical solution is derived. The derived approximate analytical solution is justified by comparing to the fusion weights/mixture weights in [8, 32]. The fusion framework can find out reliable cues and rely more on them dynamically. We test the proposed fusion framework for human tracking on a very challenging surveillance video taken at crowded subway station. We also test the fusion framework for articulated tracking. The claim that the proposed fusion framework can integrate weak modalities to improve tracking performance is supported by the promising results.

1. Introduction

Object tracking is the first step of many vision applications such as video surveillance, perceptual user interfaces, automated video content retrieval, and audio-visual speech analysis. It has drawn attentions from the researchers in vision community for at least two decades. In order to achieve robust tracking, **four** important questions have to be answered thoroughly: 1) What kind of features or similarity measures should be used, considering background clutter and distraction from other objects of similar appearance? 2) What is the searching scheme/optimization approach to find the most possible object positions? 3) How to tolerate the appearance variance of the object? By robust feature or appearance updating? 4) How do we handle the partial/total occlusion between objects?

The different elegant answers to above questions constitute a large literature in object tracking. For general objects treated as blob, e.g. face/head, car, human body, bees etc, various features have been used such as edges/contours[5, 9], texture/appearance subspace [6, 21, 17], steerable wavelets coefficients [19], gray level/color/gradient distributions [12, 28, 32], salient features [30, 10], segmented foreground pixels [40, 31], object detection/classification results [27, 2, 34, 1].

Two characters make a searching scheme/optimization approach good. First, efficient. Second, the algorithm should be capable of keeping multiple hypotheses if the ambiguity is big. However, it seems that these two desirable characters reject each other. By taking the trade off between these two characters, researchers in vision community proposed many schemes. To name a few successful and popular ones, they are: Searching schemes based on gradient descent such as, meanshift [12, 11, 37], incremental gradient method [24, 3], Newton-style method [16, 14, 1]; Searching schemes based on sampling such as, sequential Monte Carlo/particle filtering [18, 29, 32] and MCMC [26, 38]; Searching based on EM [39]; Coarse to fine searching [4]; Data association based on Dynamic Programming [23] etc.

An inborn characteristic of the vision based tracking is that the appearances of the tracking target and the background are inevitably changing, albeit gradually. Since it is extremely hard to find, if any, the general invariant features for robust tracking, quite a few successful methods handle the appearance variation by template updating or subspace learning method [6, 21, 19, 33, 10, 7].

For subspace based appearance learning algorithm [6, 21], the view based models, usually learned with Principal Component Analysis (PCA), can capture the variations in pose and lighting and can be integrated into incremen-

tal tracking framework. However, for objects of different category or even different objects of the same category, the subspace based appearance learning methods require supervised offline training, for which the labeling involves a very tedious work.

In the enlightening work of Jepson *et al*. [20], an online appearance learning algorithm is proposed. The appearance is modeled as a generative model containing 3 components: the stable component $\mathcal{S}$, the wandering component $\mathcal{W}$, and the occlusion component $\mathcal{L}$. The stable component identifying the most reliable structure for motion estimation and the wandering component representing the variation of the appearance are two Gaussian distributions. The occlusion component accounting for data outliers is uniformly distributed on the possible intensity level. Using the phase parts of the steerable wavelet coefficients [15] as feature, this algorithm achieves satisfactory tracking results. But the occlusion model $\mathcal{L}$ is a weak prior model, therefore the tracker may still lock on the background if the object is fully occluded. If this is the case, the appearance learning will take the stable background region to update the appearance model and the tracker is doomed. Their experiments verified a common fact for tracker based on appearance updating: As soon as the tracker lock on some part of the background, the tracker fails for the rest of the video sequence.

A very large percentage of tracking failures are caused by partial/total occlusion. If the target is occluded by a distracting object and the tracker uses the appearance of the distracting object to update the template, the tracker is in big trouble. Therefore, the occlusion handling is indispensable for robust tracker. Wu *et al*. [36] treat the occlusion problem as a statistical inference problem on Dynamic Bayesian Network (DBN). In [20], a uniform distributed appearance model is introduced to model the occlusion. These proposed approaches alleviate the occlusion problems to some extent, but none of existing approaches solve the partial/total occlusion completely. Reader can refer to the failure video of [20] online.

Given the above difficulties, i.e. ambiguity caused by clutter background and distracting objects, appearance variation and occlusion, it seems that none of existing features/cues alone can achieve robust tracking. In stead, successful systems have to draw from the strengths of multiple cues/modalities [35, 22].

However, the fusion of the cues from different modalities is a tough problem. What should be the optimal criterion to fuse the cues? How to adaptively weight the cues across long sequence? These questions need to be answered in principle. Current algorithm either weight multimodality cues empirically or equally after inside-modal normalization [5]. For instance, the α in [27], which is a weighting factor between dynamic model and detection results, is either set to be a fixed number or switched between some numbers according to a predefined rule.

We propose a framework to integrate multiple cues for tracking based on fusion of dynamic proposal distributions. In the setup of Bayesian sequential estimation, we propose to use the dynamic mixture of proposal distributions to substitute the state prediction probability $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. The proposal distributions are from different modalities. For example, in tracking scenario, background subtraction modal can propose a distribution of the possible blob locations. Given the position of the blob in $t-1$ frame, the dynamic model can also predict the distribution of the blob locations in t frame. Object detector as well produces a distribution of the possible object locations by densely scan the input image. By dynamically mixing these proposal distributions, we can fuse the cues from different modalities.

The fusion problem is then formulated as an optimization problem with a non-convex objective function. We convert the optimization problem on the non-convex objective function to a constrained convex programming problem. The equations to find the global optimal solution are derived based on Karush-Kuhn-Tucker theorem [25]. After carefully study the matrix structure in our specific optimization problem, we give the approximate optimal solution of the equations **analytically**.

The main contributions of this work are: 1)Formulating the problem of multimodality fusion for tracking as an optimal weighting problem among dynamic proposal distributions; 2) Converting the optimization of the non-convex objective function to a convex programming problem and give the analytical solution; 3) Giving a numerical approximation of the analytical solution based on sampling method.

2. Bayesian Sequential Estimation with Dynamic Proposal Distributions

Let $\mathbf{x}_t$ denote the state of the object of interest, and $\mathbf{y}^t = (\mathbf{y}_1 \dots \mathbf{y}_t)$ the observations up to time t. For tracking, the distribution of interest is the posterior distribution $p(\mathbf{x}_t|\mathbf{y}^t)$. However, the analytical representation of the posterior distribution is rarely available and it is computationally very expensive to directly sample this posterior distribution. Therefore, Bayesian sequential estimation method tries to compute this distribution using the following two step recursion:

prediction:

$$p(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(d\mathbf{x}_{t-1}|\mathbf{y}^{t-1}) \qquad (1)$$

update:

$$p(\mathbf{x}_t|\mathbf{y}^t) = \frac{L(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1})}{\int L(\mathbf{y}_t|\mathbf{x}_t)p(d\mathbf{x}_t|\mathbf{y}^{t-1})} \qquad (2)$$

The recursion requires a prediction model describing the state evolution probability $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and a model that

gives the likelihood of any state in the light of the current observation, $L(\mathbf{y}_t|\mathbf{x}_t)$.

Given the state of the object in frame $t-1$, we want to fuse the object state estimations from different cues in frame t. By model the state evolution probability $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ as a dynamic mixture of proposal distributions corresponding to estimations from different cues, we can achieve fusion of multi-modal cues:

$$p_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \sum_{i=1}^{M} \alpha_i(t) q_i(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{3}$$

where M is the number of multimodal cues available and $\sum_{i=1}^{M} \alpha_i(t) = 1$, $\alpha_i(t) \geq 0$. The final fused estimation of the state $\mathbf{x}_t$ is:

$$\begin{aligned} E[\mathbf{x}_t|\mathbf{y}^t] &= \sum_{i=1}^{M} \alpha_i(t) E_i[\mathbf{x}_t|\mathbf{y}^t] \\ &= \sum_{i=1}^{M} \alpha_i(t) \frac{\mathbf{x}_t L(\mathbf{y}_t|\mathbf{x}_t) p_i(\mathbf{x}_t|\mathbf{y}^{t-1})}{\int L(\mathbf{y}_t|\mathbf{x}_t) p_i(d\mathbf{x}_t|\mathbf{y}^{t-1})}, \end{aligned} \tag{4}$$

where $p_i(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int q_i(\mathbf{x}_t|\mathbf{x}_{t-1}) p(d\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$.

The **KEY** questions are: How to dynamically weight the estimations from different cues? What is the optimal criterion?

We want to use a linear combination of the proposal distributions to approach the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_t)$ so that the prediction model $p_t(\mathbf{x}_t|\mathbf{x}_{t-1})$ should coincide with $p(\mathbf{x}_t|\mathbf{y}_t)$, the normalized likelihood $L(\mathbf{y}_t|\mathbf{x}_t)$. In the L_2 space, which is a Hilbert space, we can bound the correlation between the likelihood and the prediction model by Cauchy-Schwartz inequality:

$$|\langle p_t, L_t \rangle| \leq \|L_t\| \|p_t\|, \tag{5}$$

where the inner product is defined in L_2 space: $\langle p_t, L_t \rangle = \int L(\mathbf{y}_t|\mathbf{x}_t) p(d\mathbf{x}_t|\mathbf{x}_{t-1})$.

Therefore we have:

$$\frac{|\langle p_t, L_t \rangle|}{\|p_t\|} \leq \|L_t\|. \tag{6}$$

Noticing that $\langle p_t, L_t \rangle \geq 0$, the problem is to find a set of $\alpha_i(t)$'s to

$$\begin{aligned} \text{minimize } & -\frac{\langle p_t, L_t \rangle}{\|p_t\|} \\ \text{subject to: } & \sum_{i=1}^{M} \alpha_i(t) = 1 \\ & -\alpha_i(t) \leq 0, \end{aligned} \tag{7}$$

i.e. to find α_t that

$$\begin{aligned} \text{minimize } & -\frac{\alpha_t^T \mathbf{b}_t}{\sqrt{\alpha_t^T \mathbf{G} \alpha_t}} \\ \text{subject to: } & \mathbf{1}^T \alpha_t = 1 \\ & -\alpha_t \leq \mathbf{0}, \end{aligned} \tag{8}$$

where $\alpha_t = \begin{pmatrix} \alpha_1(t) & \dots & \alpha_M(t) \end{pmatrix}^T$, $\mathbf{b}_t = \begin{pmatrix} \langle q_1, L_t \rangle & \dots & \langle q_M, L_t \rangle \end{pmatrix}^T$, and the Gram matrix

$$\mathbf{G} = G(q_1, q_2, \dots, q_M) = \begin{pmatrix} \langle q_1, q_1 \rangle & \dots & \langle q_1, q_M \rangle \\ \vdots & \ddots & \vdots \\ \langle q_M, q_1 \rangle & \dots & \langle q_M, q_M \rangle \end{pmatrix}.$$

3. Optimal Fusion of Dynamic Proposal Distributions

Since the objective function $f(\alpha_t) = -\frac{\alpha_t^T \mathbf{b}_t}{\sqrt{\alpha_t^T \mathbf{G} \alpha_t}}$ is not a convex function, it is very hard to find the global optimal solution. We try to convert it to a convex programming problem. Noticing that $f(\alpha_t) = f(c\alpha_t)$, where $c > 0$ is a constant, minimizing (8) is therefore equivalent to

$$\begin{aligned} \text{minimize } & -\frac{\alpha_t^T \mathbf{b}_t}{\sqrt{\alpha_t^T \mathbf{G} \alpha_t}} \\ \text{subject to: } & -\alpha_t \leq \mathbf{0}, \end{aligned} \tag{9}$$

We only need to normalize α_t afterwards.

Since $f(\alpha_t) = f(\frac{\alpha_t}{\|\alpha_t\|})$, we can therefore convert minimizing (9) to a convex programming problem:

$$\begin{aligned} \text{minimize } & -\alpha_t^T \mathbf{b}_t \\ \text{subject to: } & \alpha_t^T \mathbf{G} \alpha_t - 1 \leq 0; \\ & -\alpha_t \leq \mathbf{0}. \end{aligned} \tag{10}$$

3.1. Optimal Solution Given by Karush-Kuhn-Tucker Theorem

The objective function in (10) is minimized at the stationary point of the Lagrangian: $-\alpha_t^T \mathbf{b}_t + \lambda(\alpha_t^T \mathbf{G} \alpha_t - 1) - \xi^T \alpha_t$, where the scalar λ and $M \times 1$ vector ξ are Lagrange multipliers.

By Karush-Kuhn-Tucker theorem, the optimal solution of α_t that minimize (10) is the solution of following equations:

$$\begin{cases} \alpha_t = \frac{1}{2\lambda} \mathbf{G}^{-1}(\mathbf{b}_t + \xi) \\ \xi \odot \alpha_t = \mathbf{0} \end{cases}, \tag{11}$$

where $\odot$ is the element wise production (E.g. $\begin{pmatrix} a & b \end{pmatrix}^T \odot \begin{pmatrix} c & d \end{pmatrix}^T = \begin{pmatrix} ac & bd \end{pmatrix}^T$). The above

equations are hard to be solved analytically for arbitrary $\mathbf{G}$ and $\mathbf{b}$. By studying the structure of $\mathbf{G}$ and $\mathbf{b}$ in our specific problem, we found a much simpler procedure, which gives the approximate optimal solution **analytically**. We first drop the constraint $-\alpha_{\mathbf{t}} \leq \mathbf{0}$ and then we show that given the special structure of the Gram matrix $\mathbf{G}$, the optimal solution $\tilde{\alpha_{\mathbf{t}}}$ that minimizes the objective function has to be nonnegative. By this way, we derive an approximate analytical solution.

3.2. Approximate Optimal Solution in Analytical Form

In order to make equations (11) analytically solvable, we first relax the constraint $-\alpha_{\mathbf{t}} \leq \mathbf{0}$. The Lagrangian then becomes: $-\alpha_t^T \mathbf{b}_t + \lambda(\alpha_t^T \mathbf{G} \alpha_t - 1)$

The optimal solution of α_t that minimize (10) without the constraint $-\alpha_{\mathbf{t}} \leq \mathbf{0}$ is:

$$\tilde{\alpha}_t = \frac{1}{2\lambda} \mathbf{G}^{-1} \mathbf{b}_t, \tag{12}$$

where $\lambda \geq 0$ by generalized Kuhn-Tucker theorem (refer to page 249 of [25]). We argue that in our specific problem, where $\mathbf{b}_t = \begin{pmatrix} \langle q_1, L_t \rangle & \dots & \langle q_M, L_t \rangle \end{pmatrix}^T \geq \mathbf{0}$, and the Gram matrix $\mathbf{G}$ is a highly diagonal dominating matrix, i.e. $G_{ii} = \langle q_i, q_i \rangle \gg G_{ij} = \langle q_i, q_j \rangle \geq 0$ for $\forall i \neq j$. Therefore, $\mathbf{G}^{-1} \approx \text{diag}\left\{ \frac{1}{\langle q_i, q_i \rangle} \right\}$.

Then we have the approximate optimal solution:

$$\tilde{\alpha}_t \approx c \begin{pmatrix} \frac{\langle q_1, L_t \rangle}{\langle q_1, q_1 \rangle} \\ \vdots \\ \frac{\langle q_M, L_t \rangle}{\langle q_M, q_M \rangle} \end{pmatrix}, \tag{13}$$

where c is a normalization constant to ensure $\sum_{i=1}^{M} \tilde{\alpha}_i(t)$=1. The approximate optimal solution always satisfy the constraint $-\tilde{\alpha}_t \leq 0$. We therefore find the approximate optimal solution of equations (11) **analytically**. With this analytical solution, we can efficiently implement the fusion algorithm based on finding the optimal dynamic weights of the proposal distributions.

3.3. Solution Discussion

The structure of the analytical solution in (13) is simple, elegant and intuitively convincing. The dynamic weight of each proposal distribution is proportional to its inner product with the likelihood function. That means we rely on the fusion results more on the proposal distributions which are similar to the likelihood function.

Each weight is also inversely proportional to the square of the L_2 norm of the corresponding proposal distribution. That means the spiky proposal distributions are downweighted. This kind of proposal distribution is a cause of the die-off particles in condensation/particle filtering and the traditional remedy based on resampling/clustering [18, 32] requires both empirical tuning and extra computation time. By downweighting this kind of distribution, we can avoid the resampling procedure as long as one of the proposal distributions from certain cue is not spiky and gives reasonable hints.

We also want to emphasize that spiky proposal distribution may give very good object localization but usually not stable enough. By the optimal weighting of the proposal distributions from multiple cues as in (13), we can achieve robust and precise tracking. Please refer to the articulated tracking experiment in section 5 for details.

Chen and Rui give an empirical "reliability" estimation in equation (25) of [8], to fuse the tracking results from vision and audio sensors and obtain impressive results. Interestingly, equation (25) of [8] is equivalent to:

$$\hat{\alpha}_t = c \begin{pmatrix} \langle q_1, L_t \rangle \\ \vdots \\ \langle q_M, L_t \rangle \end{pmatrix}, \tag{14}$$

Chen and Rui notice that the proposal distributions similar to the likelihood/posterior is more reliable and therefore should be given more weights in fusion as the conclusion we draw above through analyzing equation (13). The authors of [32] also realize the fact when they compute the mixture weights for mixture of particle filtering (equation (4) of [32]). The results in [8, 32] further confirm that the theoretically justified solution (13) also gets support from practical work.

But both of these two works neglect one fact: **Spiky proposal distribution suggesting some states with mediocre likelihood should be downweighted.** This fact is taken into consideration by our fusion framework in equation (13), i.e. we also weight the proposal distributions according to their own L_2 norm. Without this downweighting according to L_2 norm of proposal distributions, the spiky distributions may become dominating and require resampling from time to time to avoid die-off of particles.

4. Numerical Implementation with Particle Approximation

For most of tracking/sequential estimation problem, the analytical form of the likelihood function $L(\mathbf{y}_t|\mathbf{x}_t)$ is unknown, therefore we cannot directly apply equation (13) to fuse the estimation from different proposal distribution. We use the condensation/particle filtering setup to numerically realize the fusion of multiple cues. Suppose M is the number of cues/components to be fused. $q_i(\mathbf{X}_t)$ is the proposal distribution from ith cue/component. We use N_i particles to sample q_i, and we get the particles: $\mathcal{X}_i = \{\mathbf{x}_t^{(n,i)}\}_{n=1}^{N_i}$

and the particle weights $\mathcal{W}_i = \{w_t^{(n,i)}\}_{n=1}^{N_i}$. We can then compute the unnormalized optimal weights by particle approximation according to equation (13):

$$\hat{\alpha}_i(t) = \frac{\sum_{n=1}^{N_i} w_t^{(n,i)} L(\mathbf{y}_t|\mathbf{x}_t^{(n,i)})}{\langle q_i, q_i \rangle \sum_{n=1}^{N_i} w_t^{(n,i)}}. \quad (15)$$

Here $\langle q_i, q_i \rangle$ depends on the nature of the proposal distributions and can be pre-computed without examining the particle weight. For example, in tracking scenario, if one proposal distribution is foreground segmentation results, i.e. any location in foreground is equally sampled, then $\langle q_i, q_i \rangle = 1/S$, where S is the foreground area. If the proposal distribution is Gaussian with variance σ^2, $\langle q_i, q_i \rangle = \frac{1}{2\sigma\sqrt{\pi}}$.

The normalized optimal weight is computed by:

$$\tilde{\alpha}_i(t) = \frac{\hat{\alpha}_i(t)}{\sum_{i=1}^{M} \hat{\alpha}_i(t)}, \quad (16)$$

In order to give a good approximation to $L(\mathbf{y}_t|\mathbf{x}_t)$ and $q_i(\mathbf{X}_t)$, we need to draw enough samples from all $q_i(\mathbf{X}_t)$'s. However, the number of particles we can use is restricted by the computation power. Therefore, we need to distribute the particles for sampling each q_i according to there reliability history, i.e. use more particles to sample the q_is, which have large α_i in previous frame. By this way, we give more precise numerical approximation to more reliable proposal distributions.

Finally, we can use the weights computed by equation (16) to give a fused estimation of the object state in frame t according to equation (4).

5. Experimental Results

5.1. Multiple People Tracking in a VERY Challenging Scenario

We first tested our fusion framework by tracking humans in a **very** challenging subway video sequence shot for ETISEO 2006 evaluation. Like most of the subway stations in the world, the scene are very crowded with illumination changing drastically. The video of lousy quality is taken by a cheap surveillance camera. In one word, we are pushing the fusion framework to work in a general surveillance setup.

The multiple trackers are initialized automatically by a state of the art human detector based on HOG feature [13]. The state of the tracked pedestrians is (x, y, s), i.e. the (x, y) location and the scaling factor. The proposal distributions are from 3 modalities including: motion dynamics, human detection, and foreground segmentation. Particles from motion dynamics are drawn from a Gaussian kernel centered at the position predicted from $t-1$ frame. Particles from human detector are sampled from mixture of Gaussian kernels centered at detection locations. Mixture weights of the Gaussian kernels are the normalized detection scores given by HOG detector. Particles from foreground segmentation are drawn uniformly on the foreground.

The likelihood function is the combination of SSD (Sum of Squared Distance) of template matching after local histogram equalization and Bhattacharya coefficient in Lab space. By this way, we take both texture and color information into consideration. Texture information is good at determining scale s and localize object, while color histogram improves the robustness. From our experience, the Bhattacharya coefficient in Lab space gives much better results than the ones in HSV or RGB, normalized RGB spaces, especially when the video quality is below studio level.

Sample frames of tracking results achieved by fusion, human detection, and foreground segmentation, are shown in figure 1. Each tracking instance is assigned a unique ID and drawn in different color. This multiple objects tracking and identification task is much more difficult than tracking a single object. Due to false alarms of detection and heavy occlusions, we did a post-processing to remove some tracking trajectories which either have too short temporal length or could not be verified by human detector. The fusion weights for the three modalities are shown in figure 2 for tracking instance with ID **1**. The weights for the 3 modality are: α_1**: motion dynamics;** α_2**: human detector;** α_3**: foreground segmentation.**

At frame 167, the human detector gives a good localization and has strong confidence (confidence score= 2.69. In our extensive experiments on HOG detector, we have never seen a false alarm with a confidence score of 2.69.). The fusion weight computed is $\alpha_2 = 0.69$ for human detector.

At frame 445, the localization from detector for the tracked human is poor, while the tracked human moves very smoothly and is predicted very well by our linear motion model. Therefore the good fusion should rely more on motion dynamics. The weight given by the proposed method is $\alpha_1 = 0.8$ for motion dynamics, which strongly support the reasoning above.

At frame 509, detector gives several false alarms. Foreground segmentation includes large area of background due to the shadow of the subject. The weight $\alpha_1 = 0.95$ given by our fusion framework strongly suggests using motion dynamics for tracking in this frame.

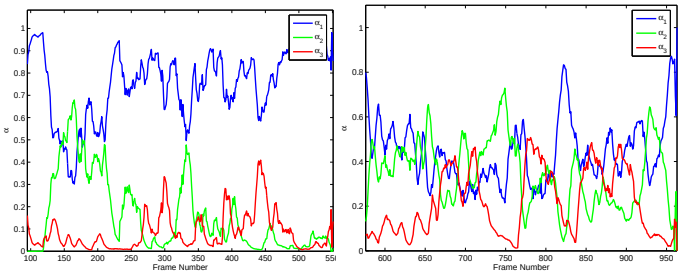
At frame 739, the segmented foreground contains big shadow area. The subject is wandering around, which fools the linear motion model. The detector correctly detects the subject with good localization in this frame. The weight is computed as $\alpha_2 = 0.73$, which suggests to rely on detector more.

At frame 822, the detector missed the subject due to heavy occlusion by another subject. The foreground also

Figure 1. **Results of multiple human tracking based on fusion the cues from motion dynamic, human detection and foreground segmentation. (From left to right, the frame numbers are:** 167**,** 445**,** 460**,** 509**,** 647**,** 739**,** 822**,** 1007 (a): the fusion tracking results. (b): human detection results (c): foreground segmentation results.

contains large area from the other subjects. Therefore the weight computed as $\alpha_1 = 0.81$ to avoid the distraction from miss detection and poor foreground segmentation.

Due to the difficulties mentioned above, our tracking results are far from perfect. But the purpose of the experiment is to demonstrate the power of the fusion framework. By comparing the results from human detection and foreground segmentation, we can draw the conclusion that the fusion framework does improve the tracking performance.



(a) ID 1 (b) ID 1

Figure 2. Dynamic fusion weights for tracking instance with ID1.

In summary, the fusion framework dynamically weight among the 3 modalities to rely on more robust modal as shown in figure 2.

5.2. Articulated Tracking Experiment

This experiment demonstrates that our fusion framework can robustly track articulated body in outdoor scenarios. In this experiment, we construct a 3D human body model with each body part represented by a truncated cone except the head. The state of the human body model is a 12 dimensional vector (global position (x, y) and 10 joint angles, two for each arm and 3 for each leg). The tracking task is to recover the state vectors from the video sequence. An object function is defined to compute the observation likelihood given the state of the human model. First we project the human body model with given state to the image plane and compute the area S of the model projection overlapping with the foreground (after background subtraction). The edges of the model projection is fitted into the edge map of the original image by the algorithm in [18] and the fitting error E is calculated. The object function takes S and E as input to summarize both region and edge information and output the observation likelihood.

We focus on periodic human body movements so that a general motion model can be learned from a large amount of video data. The motion model is simply represented by a sequence of average state and the corresponding variance.

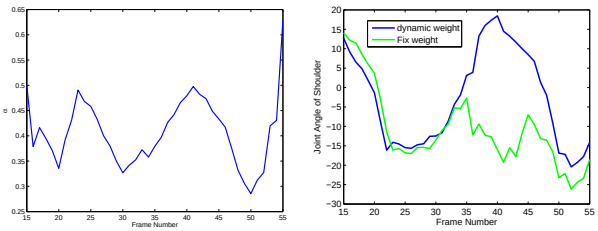
The tracker predicts samples from two proposal distributions. The first one, denoted as $\mathcal{T}$ (Temporal Motion Model), is Gaussian diffusion of the recovered state in previous frame. The other proposal distribution, denoted as $\mathcal{P}$ (Prior Motion Model), is based on the previous learned motion model. It samples from all states in the entire prior motion model with high probability at the current cycle position. The cycle position of a state is the position where the state best fit to the prior motion model and the current cycle position can be roughly computed by correlating the historical tracking results with the motion model.

The $\mathcal{P}$ proposal distribution is complementary to $\mathcal{T}$ and can help the tracker to recover the state of body parts under the hard conditions such as self-occlusion, fast motion blur or similar textures, which is extremely hart if $\mathcal{T}$ is used alone. However, the tracker may fail to recover the subtle difference between the current motion and the prior motion if it relies too much on the $\mathcal{P}$ proposal distribution. Many factors may affect the trade-off between them, for example, the closeness of the human motion to the model, the quality of the video quality, and the foreground segmentation. The

Figure 3. **Articulated tracking results: compared with the fusion by fixed weight.** (a): Results of fusion framework with dynamic fusion weights (The dynamic fusion weights are shown in figure 4(a)); (b): Results of empirical fusion with fixed weights $\alpha = 0.1$.

tracker has automatic initialization module that finds the initial cycle position by correlating a short sequence of foreground object with a cycle of template motion. The weight for each proposal distribution is then dynamically updated. Figure 3 (a) shows some sample frames of the articulated tracking results output by proposed fusion framework with dynamic α as the weight for prior motion model $\mathcal{P}$. Figure 3 (b) shows some sample frames output by the empirical fusion framework with fixed $\alpha = 0.1$. Figure 4 (a) shows the dynamic weight α for the prior motion model. Figure 4 (b) shows the joint angles of left shoulder estimated by the tracker with dynamic fusion weights and the tracker with fixed fusion weights ($\alpha = 0.1$). We can clearly see that the left arm waving pattern with the cycle around 30 frames from the joint angles estimated by dynamic fusion. We cannot see this pattern from the curve of the fixed fusion.

From figure 3 we can find that in the last two sample frames, the tracking for left arm is lost and cannot recover with the fixed fusion weight $\alpha = 0.1$, while the dynamic fusion tracker can robustly track the articulation for the whole sequence. The parameter α in this articulated tracking experiment affects the tracking performance a lot and is very hard to tune because the good choice of the value varies a lot for different test sequence. For this walking sequence, tracker with fixed fusion weights lose track at certain frames for most of the preset α's. Our fusion framework automatically find an optimal and dynamic α, which make the articulated tracking robust.



(a) (b)

Figure 4. Fusion for articulated body tracking. (a) Dynamic fusion weight α. (b) joint angle of left shoulder.

6. Conclusion

We propose a fusion framework to integrate multiple cues for tracking by finding a set of optimal dynamic weights for different tracking modalities. The optimal criterion to find the dynamic weight for each modality is given and an approximate analytical solution is derived. The derived approximate analytical solution is further justified by the work of [8, 32]. Future work is to integrate online learning component to the fusion framework to achieve robust tracking.

Acknowledgment

This work was supported in part by US Government VACE Program, and in part by Leonard Woods Institute.

References

[1] S. Avidan. Support vector tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1064–1072, 2004. 1

[2] S. Avidan. Ensemble tracking. In *CVPR(2)*, pages 494–501, Washington, DC, USA, 2005. 1

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56:221–255, 2004. 1

[4] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV*, pages 237–252, 1992. 1

[5] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232–237, 1998. 1, 2

[6] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, 26:63–84, 1998. 1

[7] J. Chen and Q. Ji. Online spatial-temporal data fusion for robust adaptive tracking. In *IEEE workshop on Online Learning for Classification in conjunction with CVPR*, 2007. 1

[8] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92(3):485 – 494, March 2004. 1, 4, 7

[9] Y. Chen, Y. Rui, and T. S. Huang. Jpdaf based hmm or real-time contour tracking. In *CVPR (1)*, pages 543–550, 2001. 1

[10] R. Collins, Y. Liu, and M. Leordeanu. On-line selection of discriminative tracking features. *IEEE Trans. PAMI*, 27(10):1631 – 1643, October 2005. 1

[11] R. T. Collins. Mean-shift blob tracking through scale space. In *CVPR*, volume 2, pages 234–240, 2003. 1

[12] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. CVPR*, pages 142–151, 2000. 1

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, Washington, DC, USA, 2005. 5

[14] Z. Fan, Y. Wu, and M. Yang. Multiple collaborative kernel tracking. In *CVPR(2)*, pages 502–509, Washington, DC, USA, 2005. 1

[15] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. PAMI*, 13(9):891–906, 1991. 2

[16] G. D. Hager, M. Dewan, and C. V. Stewart. Multiple kernel tracking with ssd. In *CVPR (1)*, pages 790–797, 2004. 1

[17] J. Ho, K.-C. Lee, M.-H. Yang, and D. J. Kriegman. Visual tracking using learned linear subspaces. In *CVPR (1)*, pages 782–789, 2004. 1

[18] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998. 1, 4, 6

[19] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *CVPR*, 2001. 1

[20] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on PAMI*, 25:1296–1311, October 2003. 2

[21] Z. Khan, T. R. Balch, and F. Dellaert. A rao-blackwellized particle filter for eigentracking. In *CVPR (2)*, pages 980–986, 2004. 1

[22] I. Leichter, M. Lindenbaum, and E. Rivlin. A general framework for combining visual trackers — the "black boxes" approach. *Int. J. Comput. Vision*, 67(3):343–363, 2006. 2

[23] M. Liu, T. X. Han, and T. S. Huang. Online appearance learning by template prediction. In *AVSS*, 2005. 1

[24] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *IJCAI*, pages 674–679, 1981. 1

[25] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, NY, USA, 1997. 2, 4

[26] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *IEEE International Conference on Decision and Control (CDC)*, 2004. 1

[27] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV (1)*, pages 28–39, 2004. 1, 2

[28] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, pages 661–675, 2002. 1

[29] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004. 1

[30] J. Shi and C. Tomasi. Good features to track. *Proc. CVPR*, pages 593–600, 1994. 1

[31] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, pages 2246–2252, 1999. 1

[32] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multimodality through mixture tracking. In *ICCV*, page 1110, 2003. 1, 4, 7

[33] J. Vermaak, P. Perez, M. Gangnet, and A. Blake. Towards improved observation models for visual tracking: selective adaptation. In *ECCV*, volume 1, pages 645–660, 2002. 1

[34] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR*, pages 951–958, 2006. 1

[35] Y. Wu and T. S. Huang. Robust visual tracking by integrating multiple cues based on co-inference learning. *Int. J. Comput. Vision*, 58(1):55–71, 2004. 2

[36] Y. Wu, T. Yu, and G. Hua. Tracking appearances with occlusions. In *CVPR (1)*, pages 789–795, 2003. 2

[37] C. Yang, R. Duraiswami, and L. S. Davis. Efficient mean-shift tracking via a new similarity measure. In *CVPR (1)*, pages 176–183, 2005. 1

[38] Q. Yu, I. Cohen, G. Medioni, and B. Wu. Boosted markov chain monte carlo data association for multiple target detection and tracking. In *ICPR*, pages 675–678, 2006. 1

[39] T. Yu and Y. Wu. Differential tracking based on spatial-appearance model (sam). In *CVPR*, pages 720–727, 2006. 1

[40] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Trans. PAMI*, 26(9):1208–1221, 2004. 1