Particle filter-based visual tracking by fusing multiple cues with context-sensitive reliabilities

Suivi d'objet dans des séquences d'images par filtrage particulaire et fusion d'informations, prenant en compte leur fiabilité

Erkut Erdem
Séverine Dubuisson
Isabelle Bloch

2010D002

janvier 2010

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

Particle Filter-Based Visual Tracking by Fusing Multiple Cues with Context-Sensitive Reliabilities

Suivi d'objet dans des séquences d'images par filtrage particulaire et fusion d'informations, prenant en compte leur fiabilité

Erkut Erdem, Séverine Dubuisson and Isabelle Bloch

erdem@tsi.enst.fr, severine.dubuisson@lip6.fr, isabelle.bloch@telecom-paristech.fr

Abstract

Many researchers argue that fusing multiple cues increases the reliability and robustness of visual tracking. However, how the multi-cue integration is realized during tracking is still an open issue. In this work, we present a novel data fusion approach for multi-cue tracking using particle filter as the underlying framework. Our method differs from previous approaches in the following ways. First, we carry out the integration of cues both in making predictions about the object to be tracked and in verifying them through observations. Our second and more significant contribution is that both stages of integration directly depend on the dynamically-changing reliabilities of the visual cues. These two aspects of our method allow the tracker to easily adapt itself to the changes in the context, and accordingly improve the tracking accuracy by resolving the ambiguities.

Résumé

Il est couramment admis que la fusion d'informations visuelles permet d'améliorer la fiabilité et la robustesse du suivi d'objets. Cependant la manière dont est faite cette fusion pendant le suivi reste une question ouverte. Dans ce rapport, nous proposons une nouvelle méthode de fusion pour le suivi par filtrage particulaire. Les principales originalités de notre approche par rapport aux méthodes existantes sont les suivantes : la fusion est intégrée à la fois dans l'étape de prédiction et dans la mise à jour en fonction des observations, et ces deux étapes de fusion prennent en compte la fiabilité de chaque information utilisée et son évolution au cours du temps en fonction du contexte. Ces caractéristiques de la méthode proposée permettent un suivi adaptatif en fonction du contexte, qui conduit à de meilleurs résultats que les approches classiques, comme le montrent les expériences menées sur plusieurs séquences d'images.

I. Introduction

Visual tracking is a widely studied topic in computer vision for a wide range of application areas. These include visual surveillance, activity analysis, man-machine interaction, augmented reality, etc. Here we consider the task of locating an object of interest on each frame of a given video sequence. This object of interest can be an actual object in the scene, e.g. a person, or a specific image region of prime importance, e.g. a face. For real-world applications, it is generally accepted that tracking based on a single visual feature would be likely to fail due to the complexity of the data and the tracking process. Thus, it has been argued in many works that considering multi-modal data leads to an improvement in tracking. It increases the robustness by letting complementary observations from different sources work together.

Erkut Erdem and Isabelle Bloch are with Institut TELECOM, Télécom ParisTech, CNRS LTCI, Paris, France.
Séverine Dubuisson is with the Laboratoire d'Informatique de Paris 6, UPMC, Paris, France.

These sources are either the visual features extracted from the same image sequence, such as color and motion cues, or the visual cues coming from different physical sensors, such as from a CCD or an infrared camera. However, how the information extracted from these sources is combined in tracking is still an open problem.

A. Related Work

Tracking methods generally involve two key processes: generating hypotheses through a prediction step and then verifying these hypotheses through some measurements. Considering the vast number of studies in tracking literature, the most general way of performing data fusion is in the measurement step. For example, in an early work [5], Birchfield suggested to combine two orthogonal visual cues (*color* and *intensity gradients*) within a hypothesize-and-test procedure. In these studies, each cue provides a likelihood or a matching score for the possible positions of the object, and the final output is determined by taking into account the product of individual likelihoods or the summation of the matching scores. The main problem with this approach is that all the modalities are given an equal reliability, which is a very unrealistic assumption. Thus, if one of visual cues becomes unreliable, it may result in a wrong estimate. There are two main approaches to overcome this issue and both of them focus on *adaptivity*.

The first group of works [21], [19], [15], [6] assigns different reliability values to different visual cues, and takes them into consideration in the measurement step. In [21], [19], the authors formulate the fusion as the weighted average of saliency maps extracted for each cue with the weights corresponding to the cues' reliabilities. Similarly, the Sequential Monte Carlo based framework proposed in [15], [6] use adaptive weights for the cues utilized in estimating the combined likelihoods. Since the reliabilities of cues are now taken into account in the computations, in this approach, the overall likelihood is more precise. On the other hand, the weakness of these studies is that the fusion is carried out only in verifying object hypotheses against observations. The multiple cues utilized are not involved in making predictions and generating hypotheses in any way. In terms of robustness, however, this is an important direction that should be pursued as well.

Indeed the second line of works [12], [22], [18], [7] concentrates on this issue and lets the multi-modal data interact with each other more explicitly throughout the tracking process. The common characteristics of these works is that the integration is also carried out in the prediction step. For instance, the ICON-DENSATION algorithm [12] uses a fixed color model specific to the object of interest to detect blobs in the current frame and uses them in the prediction step of a shape-based particle filter tracker. In [22], the authors suggested an approximate co-inference among the modalities by decoupling the object state and the measurements according to color and shape and by letting each visual cue provide hypotheses for the other one. Thus, in their formulation, the shape samples are drawn according to the color measurements, and the color samples are drawn according to the shape measurements. The tracker in [18] uses a two-layered sampling structure. The first layer constructed considering either motion or sound cues provides a coarse information regarding the object to be tracked, which is then refined by the second layer by taking account of color cues. The work in [7] also suggests a two-level, but more centralized, particle filter architecture. At the lower level, the individual trackers based on different cues perform tracking independently. At the upper level, a fuser integrates the trackers' outputs to construct more reliable hypotheses, and in return provides a feedback to the individual trackers. Although the studies that can be categorized within this latter group introduce explicit interactions between multiple cues, the way these interactions occur in each study is mainly predetermined by the global scheme/architecture considered. Furthermore, the reliabilities of the visual cues are not taken into account in the measurement steps of these studies.

B. Proposed Framework

In this paper, we present a Sequential Monte Carlo based tracking algorithm that combines multi-modal data in an original way. Our main motivation is to develop a tracking algorithm that has the properties of the two groups of works mentioned previously. That is to say, we suggest to carry out the integration of

the multiple cues in both the prediction step and in the measurement step, in estimating the likelihoods. In [16], Nickel and Stiefelhagen suggested a work in a line similar to ours by combining *Democratic Integration* [21] with two-staged layered sampling [18]. They used a predetermined layer structure with each layer being adaptive in its own. For instance, the first layer is composed of stereo cues each describing a part of the object to be tracked. However, compared to theirs, our system architecture allows interactions between multiple cues to be more dynamic and flexible.

For the prediction step, we associate each particle with a specific cue and accordingly with a specific proposal function. The crucial point is that this process is defined as an adaptive process which is governed by the dynamically-changing reliabilities of the visual cues. Thus, if one cue becomes unreliable, the tendency is to lower the total number of particles associated with it and to increase the total number of particles associated with other visual cue(s). This dynamic process improves the accuracy of the predictions since less reliable proposal functions are utilized less in the sequential importance sampling. During the prediction step no cue is given a preference over another, and the interactions between the cues are directly determined by the current context in an adaptive manner. As mentioned above, we take into account the reliabilities of the visual cues in estimating the confidence measures of the particles as well. We define the overall likelihood function so that the measurements from each cue contribute the overall likelihood according to its reliability. In return, we obtain more precise likelihood values in the measurement step as the misleading effects of the unreliable cues are reduced.

The remainder of the paper is organized as follows: Section II recalls the Sequential Monte Carlo method with a focus on multi-modal tracking. Section III gives the basis of our object model and the corresponding state dynamics. Section IV introduces the visual cues and the proposal functions that we consider in our experiments. Section V gives the outline of our multi-modal tracking algorithm and our main contributions. Section VI presents some illustrative tracking experiments in which we analyze the performance of the proposed algorithm. Finally, Section VII makes a brief summary of our work, and points out the future directions.

II. Sequential Monte Carlo and Multi-modal Tracking

In a classical filtering framework, the main aim is to estimate the posterior distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$ of the state vector $\mathbf{x}_k$ through a set of measurements $\mathbf{y}_{1:k}$ up to the current time step k. The Bayesian sequential estimation approach computes this distribution according to a two-step recursion: a *prediction* step

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1}) d\mathbf{x}_{k-1} \tag{1}$$

followed by a *filtering* step

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \propto p(\mathbf{y}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{y}_{1:k-1}). \tag{2}$$

This formulation requires two models to be defined: an evolution (transition) model for the state dynamics $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ and a likelihood model for the observations $p(\mathbf{y}_k \mid \mathbf{x}_k)$. One can obtain an optimal solution to the posterior distribution under highly restrictive assumptions. For instance, the Kalman filter [14] assumes these models to be linear and Gaussian. However, real data is generally non-linear, non-Gaussian, multi-modal in nature, necessitating the use of some approximation techniques. In this regard, since its introduction over a decade ago, Sequential Monte Carlo based filtering (*also known as* particle filter) [10], [11], [3], [9] has proved to be an effective method for visual tracking. It provides a simple yet flexible solution to optimal state estimation problems.

The main idea behind particle filter is to approximate the posterior distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$ by a weighted set of N particles $\{\mathbf{x}_k^{(i)}, w_k^{(i)}\}_{i=1}^N$ as

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta_{\mathbf{x}_k^{(i)}}(\mathbf{x}_k). \tag{3}$$

with $\delta_{\mathbf{x}_0}$ denoting the Dirac delta mass centered on x_0, and each particle representing a possible state $\mathbf{x}_k$ and its weight $w_k^{(i)} \in [0,1]$ describing its confidence measure.

The recursive estimation process is, then, characterized by two main steps: with an approximation of $p(\mathbf{x}_{k-1} \mid \mathbf{y}_{1:k-1})$ at hand new particles are generated from the old particle set $\{\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^{N}$ by making use of a known proposal function, $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k \mid \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{1:k})$. This prediction step is then followed by an update step where the weights of the new particles $w_k^{(i)}$ are determined from the new observations $\mathbf{y}_k$ using

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{y}_k \mid \mathbf{x}_k^{(i)}) p(\mathbf{x}_k^{(i)} \mid \mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k \mid \mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_{1:k})} \tag{4}$$

with $\sum_{i=1}^{N} w_k^{(i)} = 1$. As a further step, a resampling phase, which removes the particles with low weights and accumulates the particles with high weights, can be employed to avoid the degeneracy of the particles [10]. Generally, the final tracking decision is made by taking into account the conditional mean, the weighted average of the particles $\{\mathbf{x}_k^{(i)}\}$, or the particles with the highest weights.

For multi-modal tracking, the simplicity and the flexibility of the particle filter offer a wide variety of solutions. One direction is to perform data fusion in the likelihood estimation step. In this regard, the most straightforward way of integrating multiple measurement sources is to assume that these measurements are conditionally independent given the state and subsequently factorize the overall likelihood as

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{m=1}^{M} p(\mathbf{y}^m \mid \mathbf{x}) \tag{5}$$

with M representing the total number of sources. As we stated in the introduction, it is possible to increase the accuracy of the combined likelihood by further considering the reliabilities of the measurement sources in the integration phase [19], [15], [6].

The studies [12], [22], [18], [7] consider another direction and suggest explicit interactions between different modalities. In these works, the main emphasis is on the proposal functions utilized in the prediction step, and how the candidate state hypothesis proposed by different modalities can be integrated.

III. Object Model and State Dynamics

The tracking framework that we propose in this work does not depend on a specific object model, and any model suggested in literature can be utilized. However, it is important to note that the model of choice restricts the visual cues employed in the tracking process. In this paper, we prefer to use a simple model and represent the object to be tracked by a fixed reference rectangular region parameterized as $\Omega = (x^c, y^c, w, h)$, where (x^c, y^c) denote the coordinates of the center of the rectangular region having a width w and a height h.

Considering the transformation of the reference region throughout the tracking sequence, we define the object state as $\mathbf{x}_k = (x_k, y_k, s_k, t_k) \in \mathcal{X}$. It describes a new region $\Omega_{\mathbf{x}_k} = (x_k, y_k, s_k w, t_k h)$ with s_k and t_k denoting the scaling factors for the width and the height of the reference region, respectively.

For the state evolution model, we assume mutually independent Gaussian random walk models along with a small uniform component as in [18]. This uniform component is used to compensate the irregular motion behavior of the object that is tracked and provides a kind of re-initialization. Accordingly, the state evolution model can be written as:

$$p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \sim \beta_U \mathcal{U}(\mathbf{0}, \mathbf{x}_{max}) + (1 - \beta_U) \mathcal{N}(\mathbf{x}_{k-1}, \Lambda) \tag{6}$$

where $\mathcal{U}$ denotes the uniform distribution with the vector $\mathbf{x}_{max}$ representing the maximum allowed values over the set $\mathcal{X}$, $\mathcal{N}(\mathbf{x}_{k-1}, \Lambda)$ denotes the Gaussian distribution with mean $\mathbf{x}_{k-1}$ and covariance $\Lambda = diag(\sigma_x^2, \sigma_y^2, \sigma_s^2, \sigma_t^2)$, and β_U is the weight of the uniform component which is generally set to a small value. Additionally, the initial state of the object is assumed to be described by a uniform distribution $p(\mathbf{x}_0) = \mathcal{U}(\mathbf{0}, \mathbf{x}_{max})$. In all the experiments reported in this paper, we use $\sigma_x^2 = \sigma_y^2 = 3$, $\sigma_s^2 = \sigma_t^2 = 0.01$, and $\beta_U = 0.01$.

IV. Visual Cues and Proposal Functions

This section describes the visual cues that we utilize in tracking an object of interest. These are simply *color*, *motion* and *infrared brightness*, and are discussed in the following subsections in detail. Before going into details, let us first present how we employ them throughout the tracking process in a more general way.

In our work, while extracting these visual cues from an image frame, we follow a conventional approach and use measurements based on histograms. We compute the likelihoods and construct the individual proposal functions by making use of reference histograms which are defined for each visual cue. There are several strategies for designing these reference histograms. For example, they can be constructed manually by hand by taking into account the properties of the object in terms of the visual cue considered (e.g. a color model for human skin) or they can be constructed from a particular frame of the tracking sequence by letting the user specify the region of interest. We consider both strategies while constructing our reference histograms, and use these histograms throughout the whole tracking sequence without updating them.

Mainly, the construction of the proposal functions and the estimation of the likelihoods depend on the comparison between the histograms extracted from the candidate regions and the reference histogram. For that, we utilize the Bhattacharyya histogram similarity measure [4], which is defined as:

$$D(\mathbf{h}_1, \mathbf{h}_2) = \left(1 - \sum_{i=1}^{B} \sqrt{h_{i,1} h_{i,2}}\right)^{1/2} \tag{7}$$

where B denotes the number of bins, and $h_{i,1}$ represents the i^{th} bin of histogram 1.

It is important to note that, as in [18], the proposal functions described in the subsequent subsections are defined only for suggesting the new values for the location component of the object state. For the scaling factors, the proposal functions are taken as the corresponding component of the state evolution model described in Equation (6).

A. Color Cue

Color is one of the most widely used visual cues in tracking frameworks. Its widespread use is due to its characteristics that allow encoding the appearance of the object tracked in an efficient and robust way. In this work, following [17], we adopt an observation model that is based on Hue-Saturation-Value (HSV) color histograms with $B_C = B_h B_s + B_v$ bins. While we populate the first $B_h B_s$ bins with the pixels having saturation and value greater than some pre-defined thresholds (in our experiments we used 0.1 and 0.5, respectively), we include the value information in the additional B_v bins considering the remaining pixels. Using this definition, we define our color likelihood as

$$p(\mathbf{y}^C \mid \mathbf{x}) \propto \exp\left(-D^2(\mathbf{h}^C_{\mathbf{x}}, \mathbf{h}^C_{ref})/2\sigma_C^2\right) \tag{8}$$

with $\mathbf{h}^C_{ref}$ denoting the B_C-bin normalized reference histogram, $\mathbf{h}^C_{\mathbf{x}}$ representing the normalized color histogram which is obtained from a candidate object region specified by the object state $\mathbf{x}$, and $D^2(\mathbf{h}^C_{\mathbf{x}}, \mathbf{h}^C_{ref})$ being the Bhattacharyya histogram similarity measure between them.

The construction of the proposal function also depends on the color likelihood model described above. Typically, we first estimate the color likelihoods on a subset of image locations over the current frame. For this, we use a pre-defined step size of 5 pixels through the current frame, and keep the scale factors fixed as $s = t = 1$. The likelihoods estimated in this way define an approximate probability distribution map for the object tracked. Figure 1 illustrates a sample construction of these likelihoods. In fact, these approximate distribution maps, which are estimated for each visual cue, are of critical importance in estimating the reliabilities of the cues which guides the whole tracking process. The details of this procedure will be given in Section V. Once these likelihoods are estimated, we define our proposal function as follows:

$$\begin{aligned} q^C(x_k, y_k \mid x_{k-1}, y_{k-1}, \mathbf{y}^C_k) &= \beta_{RW} \mathcal{N}\left((x_{k-1}, y_{k-1}), (\sigma_x^2, \sigma_y^2)\right) \\ &+ \frac{(1-\beta_{RW})}{N_C} \sum_{i=1}^{N_C} \mathcal{N}\left(\mathbf{p}^C_i, (\sigma_x^2, \sigma_y^2)\right). \end{aligned} \tag{9}$$

(a) (b) (c)

Fig. 1. Color likelihood. (a) The frame and the region where the reference histogram is contructed. (b) A sample frame from the tracking sequence. (c) The approximate probability distribution map estimated for the frame given in (b). Observe the responses in the true location of the object of interest along with the other image regions having an appearance similar to the one of the reference region.

In Equation (9), the first component is the Gaussian random walk component for the object location that we previously introduced in our state evolution model given in Equation (6). The points $\mathbf{p}_i^C = (x_i, y_i), i = 1, \cdots, N_C$ denote the image locations having a likelihood greater than a threshold (i.e. $p(\mathbf{y}^C \mid \mathbf{x}) > \tau^C$), and define the centers of Gaussians in the mixture model utilized in the second component, respectively. Generally, the coefficient β_{RW} is set to a relatively high value (e.g., $\beta_{RW} = 0.75$), and thus the main tendency is to preserve the smoothness of the tracking trajectory. On the other hand, the second component allows jumps in the state space to the image regions that likely contain the object which is tracked.

B. Motion Cue

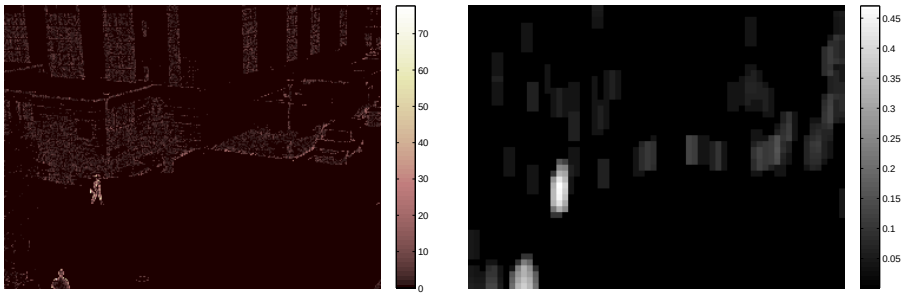
The motion activity in a tracking sequence is an important indication for the object of interest, especially when the video sequence was assumed to be captured by a static camera and the object that is tracked is generally in motion. Considering such an assumption, the image locations having a motion activity at the frame k can be determined from the absolute difference of the intensity images at the frames k and $k-1$. In the frame difference, the pixels with large values indicate the motion activity. If there is no motion, the frame difference is either zero or has a very small value due to the noise and/or due to the slight changes in the intensity.

To estimate the motion likelihood, we follow the approach suggested in [18]. For a region of interest specified by the state $\mathbf{x}$, we associate a motion histogram $\mathbf{h}_{\mathbf{x}}^M = (h_{1,\mathbf{x}}^M, \cdots, h_{B_M,\mathbf{x}}^M)$ with B_M denoting the number of bins. During populating the histogram, we enlarge the regions of interest by a few pixels (five in our experiments). This guarantees the inclusion of the silhouettes and allows capturing the motion activity across them. On the other hand, the reference histogram $\mathbf{h}_{ref}^M$ is defined considering a uniform distribution, i.e. $h_{i,ref}^M = \frac{1}{B_M}, \quad i = 1, \cdots, B_M$. The rationale behind this definition depends on the characteristics of the motion histogram extracted from the candidate regions. Typically, when there is no motion in the candidate region, only the lowest bins of the histogram are populated. For the case of motion, the candidate motion histogram shows an irregular structure. Thus, comparing it to a uniform distribution reveals information regarding the motion likelihood. In the case of no motion activity, the Bhattacharyya histogram similarity measure yields $D_{no\_motion}^2 = 1 - \sqrt{1/B_M}$. Considering this, we define the motion likelihood as

$$p(\mathbf{y}^M \mid \mathbf{x}) \propto 1 - \exp\left(-(D_{no\_motion}^2 - D^2(\mathbf{h}_{\mathbf{x}}^M, \mathbf{h}_{ref}^M))/2\sigma_M^2\right). \tag{10}$$

Figure 2 illustrates the result of this procedure for a sample frame.

As in Section IV-A, the proposal function is constructed by estimating the likelihoods on a subset of image locations over the current frame. While estimating them, again the scale factors are fixed as $s = t = 1$ and the pre-defined step size is used. The locations having a likelihood greater than a threshold ($p(\mathbf{y}^M \mid \mathbf{x}) > \tau^M$) denoted by $\mathbf{p}_i^M = (x_i, y_i), i = 1, \cdots, N_M$ are then used, as in [18], to define the



(a) (b)

Fig. 2. Motion likelihood. (a) The absolute frame difference for two successive frames. (b) The approximate probability distribution map estimated from (b). Notice that we have two strong responses which leads to an ambiguity. This is due to a second person entering the scene from bottom-left part of the frame.

proposal function as

$$\begin{aligned} q^M(x_k, y_k \mid x_{k-1}, y_{k-1}, \mathbf{y}_k^M) &= \beta_{RW}\mathcal{N}\left((x_{k-1}, y_{k-1}), (\sigma_x^2, \sigma_y^2)\right) \\ &+ \frac{(1-\beta_{RW})}{N_M}\sum_{i=1}^{N_M}\mathcal{N}\left(\mathbf{p}_i^M, (\sigma_x^2, \sigma_y^2)\right). \end{aligned} \tag{11}$$

C. Infrared Brightness Cue

Besides color and motion, we additionally employ infrared brightness cue in our experiments. This cue requires the tracking sequence to be imaged from an infrared camera, and allows us to consider different thermal characteristics of an object of interest during tracking. In estimating the likelihoods and constructing the corresponding proposal function, we follow an approach similar to the ones explained in the previous subsections. We populate our histograms using the brightness values within an infrared image region. For the reference histogram, this region can be specified by the user from a particular frame. Then, we define the infrared brightness likelihood as

$$p(\mathbf{y}^I \mid \mathbf{x}) \propto \exp\left(-D^2(\mathbf{h}_{\mathbf{x}}^I, \mathbf{h}_{ref}^I)/2\sigma_I^2\right). \tag{12}$$

where $\mathbf{h}_{ref}^I = (h_{1,ref}^I, \cdots, h_{B_I,ref}^I)$ is the B_I-bin normalized reference histogram, and $\mathbf{h}_{\mathbf{x}}^I = (h_{1,\mathbf{x}}^I, \cdots, h_{B_I,\mathbf{x}}^I)$ denotes the normalized brightness histogram obtained from the candidate object region. In Figure 3, we illustrate a sample construction of these likelihoods.

Subsequently, we construct the proposal function by estimating the likelihoods on a subset of image locations over the current frame and using the locations having a likelihood greater than a threshold, i.e. $p(\mathbf{y}^I \mid \mathbf{x}) > \tau^I$, as follows:

$$\begin{aligned} q^I(x_k, y_k \mid x_{k-1}, y_{k-1}, \mathbf{y}_k^I) &= \beta_{RW}\mathcal{N}\left((x_{k-1}, y_{k-1}), (\sigma_x^2, \sigma_y^2)\right) \\ &+ \frac{(1-\beta_{RW})}{N_I}\sum_{i=1}^{N_I}\mathcal{N}\left(\mathbf{p}_i^I, (\sigma_x^2, \sigma_y^2)\right) \end{aligned} \tag{13}$$

where $\mathbf{p}_i^I = (x_i, y_i), i = 1, \cdots, N_I$ denote the image locations where the object tracked is likely to be.

In our experiments, we fixed $\sigma_C = 0.2$, $\sigma_M = 0.4$, $\sigma_I = 0.25$, $B_h = B_s = B_v = 10$, $B_M = 20$, $B_I = 30$, and used detection rates $\tau^C = \tau^I = 0.65$, $\tau^M = 0.2$. In Equations (9), (11) and (13), if respectively N_C, N_I or N_M equals to zero, we use only the first Gaussian random walk component for the related proposal function.

(a) (b) (c)

Fig. 3. Infrared brightness likelihood. (a) The frame and the region where the reference histogram is constructed. (b) A sample frame from the tracking sequence. (c) The approximate probability distribution map estimated for the frame given in (b). Notice the ambiguities; both the object of interest and the van produce very strong responses.

V. Tracking Algorithm

We propose a novel approach for integrating different visual cues during tracking. Unlike the previous works summarized in Section I-A, we do not give preference to any cue and do not use a global scheme with a predetermined structure. We mainly let the current visual context determine how the interactions between the multiple cues are carried out. In all phases of the tracking process, we emphasize the information derived from the reliable cues and ignore the information provided by the unreliable cues. This view certainly involves discovering and using the relative reliabilities of the visual cues.

In this respect, the first and the most crucial step is adjusting cues reliabilities with respect to the current context. For that, we adopt the *Democratic Integration* method [21]. In *Democratic Integration*, the reliabilities are determined by considering the correlation among the visual cues. Simply, different cues try to reach an agreement on a joint result and they adapt themselves considering the result currently agreed on. This adaptive process is in accordance with the experimental results which conjectures that humans use adaptive strategies to integrate information provided by different cues or modalities [13], [20].

Specifically, *Democratic Integration* requires a quality measure s^{ℓ} to be defined for each cue which measures the degree of agreement between the joint result and the result the cue individually suggests. These measures are utilized to adjust the reliabilities so that the cues that are not in agreement with the joint result are suppressed and the cues that are in line with the joint result are given a higher influence in the future. In our work, we initialize the reliabilities with equal weights with their sum equal to 1 and define these quality measures over the approximate probability distribution maps which are also utilized in estimating the proposal functions (Section IV). As a result, the new reliabilities are estimated using the reliability values over the previous frame and the current observations.

The context-sensitive structure of the prediction step involves an adaptive assignment procedure. Each particle is assigned to a modality denoted by ℓ with $\ell \in \{C, I, M\}$ (C for color, I for infrared brightness, M for motion) and accordingly to a specific proposal function $q^{\ell_k}(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k^{\ell_k})$. This assignment process is defined as an adaptive process which is governed by the dynamically-changing reliabilities of the visual cues. Thus, if one cue becomes unreliable relative to other visual cues, the tendency is to lower the total number of particles associated with it and to increase the total number of particles associated with other visual cue(s). As a result, the tracking accuracy increases as less reliable proposal functions are utilized less in the sequential importance sampling in predicting the position of the object to be tracked.

Second, we define the overall likelihood function so that the measurements from a cue ℓ contribute to the overall likelihood according to its reliability r^{ℓ} as:

$$p(\mathbf{y}_k \mid \mathbf{x}_k) = \prod_{\ell \in \{C,I,M\}} p(\mathbf{y}_k^{\ell} \mid \mathbf{x}_k)^{r_k^{\ell}} \tag{14}$$

with $\sum_{\ell \in \{C,I,M\}} r^{\ell} = 1$. It should be noted that if we take the logarithm of the likelihood formula given in Equation (14), we get an expression which is in a certain sense analogous to the weighted sum used

in the *Democratic Integration* method [21] to integrate multiple cues. Since we adjust the reliability of a visual cue in accordance with the other cues considered, the assumption on the conditional independence of the measurements gets relaxed in our formulation. This results in more precise likelihood values as it reduces the misleading effects of the unreliable cues. However, an important point is that the individual likelihoods having a value estimated as zero makes the overall likelihood zero as we take the product, whether its reliability score is low or not. Thus, in our experiments, we adjust all such likelihoods values and explicitly set them to a small value like $p(\mathbf{y}^\ell \mid \mathbf{x}) = 0.001$.

The whole algorithm is summarized below:

Algorithm

In the initialization, assume the initial states to be uniformly distributed over the state space, i.e. $p(\mathbf{x}_0) = \mathcal{U}_\mathcal{X}(\mathbf{x}_0)$.

From the particle set $\{\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N$ at the time step $k-1$, determine the new particle set for time k.

- **update** reliabilities
  - **estimate** an approximate target position $\hat{\mathbf{x}}_k$ considering the reliabilities over the previous frame and the current observations as

$$\hat{\mathbf{x}}_k = \arg\max_x \left(\hat{p}(\mathbf{y}_k \mid \mathbf{x})\right) = \arg\max_x \left(\prod_{\ell \in \{C,I,M\}} \hat{p}(\mathbf{y}_k^\ell \mid \mathbf{x})^{r_{k-1}^\ell} \right) \tag{15}$$

    with $\hat{p}(\mathbf{y}_k^\ell \mid \mathbf{x})$ denoting the approximate probability distribution map estimated for the modality denoted by ℓ
  - **estimate** the quality measures for each cue as follows:

$$s_k^\ell = \begin{cases} 0 & \textbf{if } \hat{p}(\mathbf{y}_k^\ell \mid \hat{\mathbf{x}}_k) \leq \langle \hat{p}(\mathbf{y}_k^\ell \mid \mathbf{x}) \rangle \\ \hat{p}(\mathbf{y}_k^\ell \mid \hat{\mathbf{x}}_k) - \langle \hat{p}(\mathbf{y}_k^\ell \mid \mathbf{x}) \rangle & \textbf{if } \hat{p}(\mathbf{y}_k^\ell \mid \hat{\mathbf{x}}_k) > \langle \hat{p}(\mathbf{y}_k^\ell \mid \mathbf{x}) \rangle \end{cases} \tag{16}$$

    where $\langle \cdots \rangle$ denotes the average over the approximate probability distribution map
  - **update** reliabilities considering the current quality measures as follows:

$$r_k^\ell = r_{k-1}^\ell + \eta(s_k^\ell - r_{k-1}^\ell) \tag{17}$$

    with η denoting a time constant which we set to 0.1 in our experiments.
- **simulate** $\ell_k^{(i)}$:
  - **generate** a random number $\alpha \in [0, 1)$, uniformly distributed.
  - **set** $\ell_k^{(i)} = \begin{cases} C & \textbf{if } \alpha < r_k^C \\ I & \textbf{if } r_k^C \leq \alpha < r_k^C + r_k^I \\ M & \textbf{if } \alpha \geq r_k^C + r_k^I \end{cases}$
- **simulate** $\mathbf{x}_k^{(i)} \sim q^{\ell_k^{(i)}}(\mathbf{x}_k \mid \mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k^{\ell_k^{(i)}})$
- **update** weights $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{y}_k|\mathbf{x}_k^{(i)}) p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q^{\ell_k^{(i)}}(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{y}_k^{\ell_k^{(i)}})}$ with $\sum_{i=1}^N w_k^{(i)} = 1$
- **resample**: simulate $a_i \sim \{w_k^{(n)}\}_{n=1}^N$, and replace $\{\mathbf{x}_k^{(i)}, w_k^{(i)}\} \leftarrow \{\mathbf{x}_k^{(a_i)}, \frac{1}{N}\}$
- **decision**: use the conditional mean or the particles with the highest weights

VI. Experimental Results

In this section, we demonstrate the performance of the proposed framework on illustrative video sequences. We typically compare our results obtained considering multiple cues with context-sensitive reliabilities with those obtained using a single cue or multiple cues with fixed reliabilities. For two of the illustrative sequences, we further provide the tracking outcomes of the two-layered partitioned sampling approach suggested in [18] (we use the proposal functions and the likelihoods defined in Sec. IV in

our implementation). As we mentioned in the introduction, this approach uses a two-layered partitioned sampling method in which the location of the object of interest is initially determined by using the motion cues, and subsequently refined in accordance with the color cues.

Throughout tracking, we use a fairly small number of particles, $N = 100$, and employ the conditional mean and the particles with the five highest weights to depict the outcomes. We associate different colors for the particles, and the rectangular regions they represent, depending on the cue they are attached to: *green* for color, *blue* for motion, and *red* for infrared brightness. In addition, we draw the rectangle represented by the conditional mean in *white*.

We first consider a sequence from the BEHAVE Interactions Test Case Scenarios [1] where we try to track a person with a white shirt using color and motion information. The reference color model is constructed from the rectangular region shown in Figure 4(a). Throughout the sequence, first, a group of people goes after the person of interest and attacks him. During this time, he is completely occluded. Next, at some point, the person of interest kneels down and stops moving. These different phenomena observed throughout the video sequence exemplifies the contextual changes that we exploit in our tracking framework.

As Figures 5 and 6 respectively demonstrate, the color-based tracking and the motion-based tracking may lead to inaccurate results due to the ambiguities inherent to the processing of the video sequence considering single modalities. There are objects in the background which have similar appearances to the object of interest. Therefore, soon after the initialization, the framework based on color starts tracking the wrong object and remains at this local minimum point during nearly half of the video sequence. However, it is eventually able to recover the actual object of interest with the utility of the color-based proposal. The outcomes of the motion-based tracker is much worse since the video sequence involves several persons in motion. This generally makes the motion-based proposal generate particles that do not correspond to the actual person of interest.

As one expects, considering color and motion cues all together with fixed values for reliabilities gives better tracking results than using only one modality (Figure 7). Yet, such a scheme has some drawbacks. Since equal weights are given for color and motion cues, if one of the sources becomes unreliable, it directly affects the results. In the video sequence, the person entering the scene during which the actual person of interest is at rest distracts tracking.

As illustrated in Figure 8, considering a scheme with context-sensitive reliabilities eliminates most of the ambiguities mentioned and results in an improvement in the outcomes. For instance, when the person to be tracked is occluded by the group of people following him, the reliability of the color cue decreases, and thus the motion cue particularly guides the tracking process during this time interval. Similarly, when the person of interest becomes idle, the reliability of motion decreases, making the color cue the dominant cue. Thus, the tracking process does not get distracted by the person entering the scene unlike in the case with fixed reliabilities. In Figure 9, we provide color and motion likelihoods as well as their combinations with two different strategies for a sample time instant (for the frame where the person of interest is at rest). Moreover, the changes in the reliabilities of the cues are illustrated in the plot given in Figure 10.

In Figure 11, we demonstrate the disadvantage of using a predetermined layered sampling approach by considering the global scheme proposed by Perez et al. [18]. As it can be clearly seen from the figure,

(a) (b) (c)

Fig. 4. The reference regions utilized in tracking (a) **seq. #1**, (b) **seq. #2**, and (c) **seq. #3**.

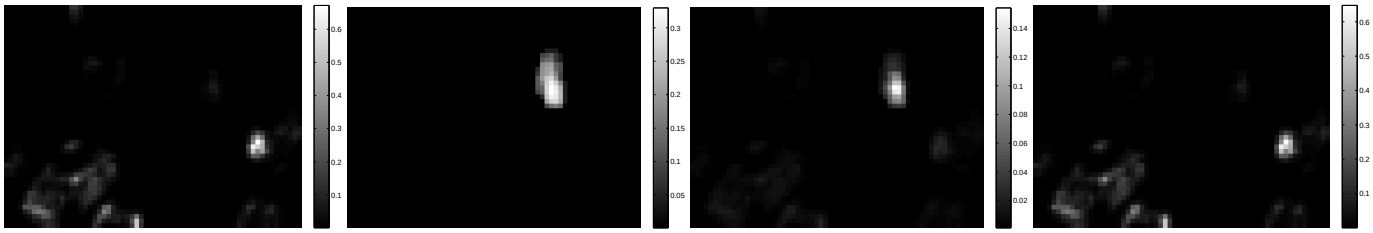
Fig. 5. **seq. #1** Sample tracking results based on color. The background contains objects having similar appearances to the person of interest. The tracker starts tracking the wrong object soon after the initialization, but eventually is able to recover the actual object of interest with the use of the color-based proposal.

Fig. 6. **seq. #1** Sample tracking results based on motion. The sequence contains several objects in motion, and thus during the tracking process the particles are distributed all around these objects instead of the actual object of interest.

Fig. 7. **seq. #1** Sample tracking results based on both color and motion with fixed reliabilities. The results are better than those that are obtained by considering single visual cues. However, this time a person entering the scene during which the actual person of interest is at rest results in inaccurate tracking.

Fig. 8. **seq. #1** Sample tracking results based on both color and motion with context-sensitive reliabilities. Modifying the reliabilities of the visual cues according to the context and using them accordingly eliminates most of the ambiguities that the previous cases (Figures 5-7) cannot easily cope with. For example, the person of interest being at rest makes the reliability of the motion cue decrease, letting the color cue be the key cue in tracking. As a result, the tracking process does not get distracted by the person entering the scene.

for the video sequence under consideration, the sampling strategy suggested in [18] results in inaccurate tracking. The tracking process relies primarily on the motion information in the prediction step, and thus the person entering the scene during the time the actual person of interest is at rest distracts the tracking process as in the case with fixed reliabilities (Figure 7). Since this approach does not attach the particles to any particular modality, we use a different color (yellow) for the particles representing the tracking



color motion combined (fixed weights) combined (adaptive weights)

Fig. 9. **seq. #1** Color and motion likelihoods, and their combinations with fixed and context-sensitive weights. Observe how the combined likelihood changes when adaptive weights for the reliabilities are considered.

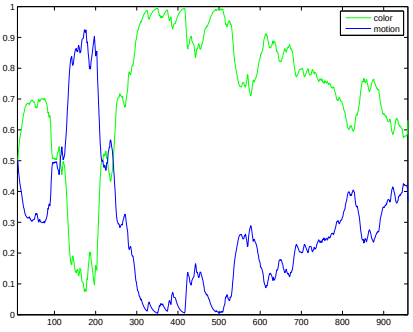


Fig. 10. **seq. #1** Reliabilities throughout the sequence.

Fig. 11. **seq. #1** Sample tracking results based on the two-layered partitioned sampling approach suggested by Perez et al. [18]. Tracking relies primarily on the motion information for the localization, and thus a person coming into the scene during the time the actual person of interest becomes idle leads to inaccurate tracking.

outcomes.

At this point, we should mention that the detection thresholds utilized are of critical importance for the proposal functions, and thus the results obtained here. For example, increasing the value of τ^M to a convenient value makes both the framework that uses fixed reliabilities for color and motion, and the two-layered partitioned sampling approach [18] accurately track the person of interest. This highlights that our proposed work is more robust against the values chosen for the detection parameters in terms of the false positives given the current context.

In the second experiment, we consider a tracking sequence captured from an infrared camera along with a CCD camera, taken from the OSU Color-Thermal Database [8]. This allows us to employ infrared brightness as another source of information during tracking. We test our framework under four scenarios: employing color and motion cues together, and using infrared brightness along with them, with and without context-sensitive reliabilities. For the color and the infrared brightness models, we use the reference regions given in Figure 4(b).

We show the results obtained by using fixed and adaptive weights for the cues' reliabilities in Figures 12 and 13, respectively. In each figure, we provide the outcomes based on color and motion, and color, motion and infrared brightness side by side. It can be seen from these figures that the results of the

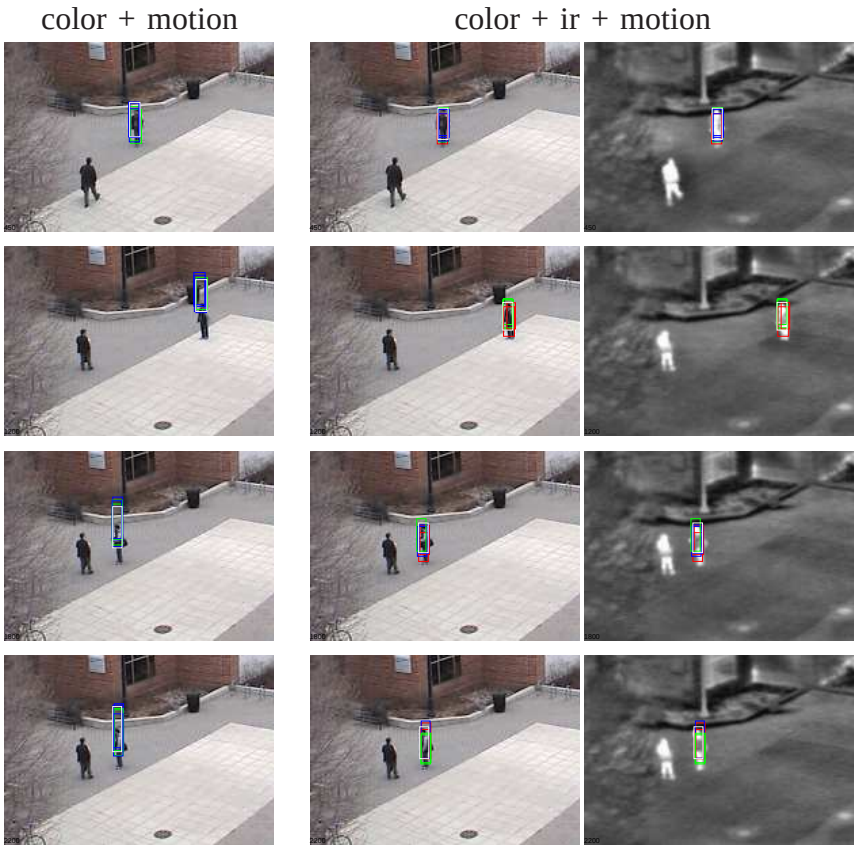


Fig. 12. **seq. #2** Sample tracking results with fixed reliabilities. The scene contains several objects that have similar appearances to the person of interest, and the person's view changes throughout the sequence. These make the reference color model quickly become inadequate. Thus, the framework built upon color and motion cues leads to enlarged and inaccurate object regions. Considering all available modalities (color, infrared brightness, and motion) improves the results but in a certain extent as the reliabilities are held constant during the tracking process.

framework built upon color and motion are not good, whether fixed values for the reliabilities are used or not. These cues both fail to account for the uncertainties in the tracking sequence. Specifically, the reference color model quickly becomes inadequate for describing the appearance of the person of interest, leading to enlarged and inaccurate object regions. This is mainly due to the changes in the person's view throughout the sequence and the nearby objects with a similar color. Introducing infrared brightness as a complementary cue, in this respect, improves the performance and provides more accurate tracking. However, it is important to note that refining the reliabilities in respect of the contextual information gives more accurate results than using fixed values for the reliabilities most of the time as infrared brightness is given a higher weight, or importance, than the other visual cues during tracking (Figure 14).

Lastly, we consider an image sequence from the CAVIAR project [2]. It consists of several people moving across the hallway in a mall, and we try to track the person specified in Figure 4(c) throughout this sequence. We again compare the tracking outcomes obtained by using single visual cues, color and motion, with that of obtained by combining these two. For color-based tracking, we construct our reference color histogram by using the rectangular region shown in Figure 4(c).

As illustrated in Figure 15, using motion data alone leads to inaccurate tracking. The sequence contains several persons moving across the hallway. The tracking process cannot distinguish the actual person of interest from the others, and the particles are distributed all over the moving persons. On the other hand, the color-based tracking and our framework provide nearly similar tracking results (Figures 16 and 17). They succeed in tracking the object for most part of the sequence, but they lose the track whenever a person having a similar appearance enters the scene. The reason behind the similar performance is that with respect to the contextual information, color is determined to be the main cue and is given a much

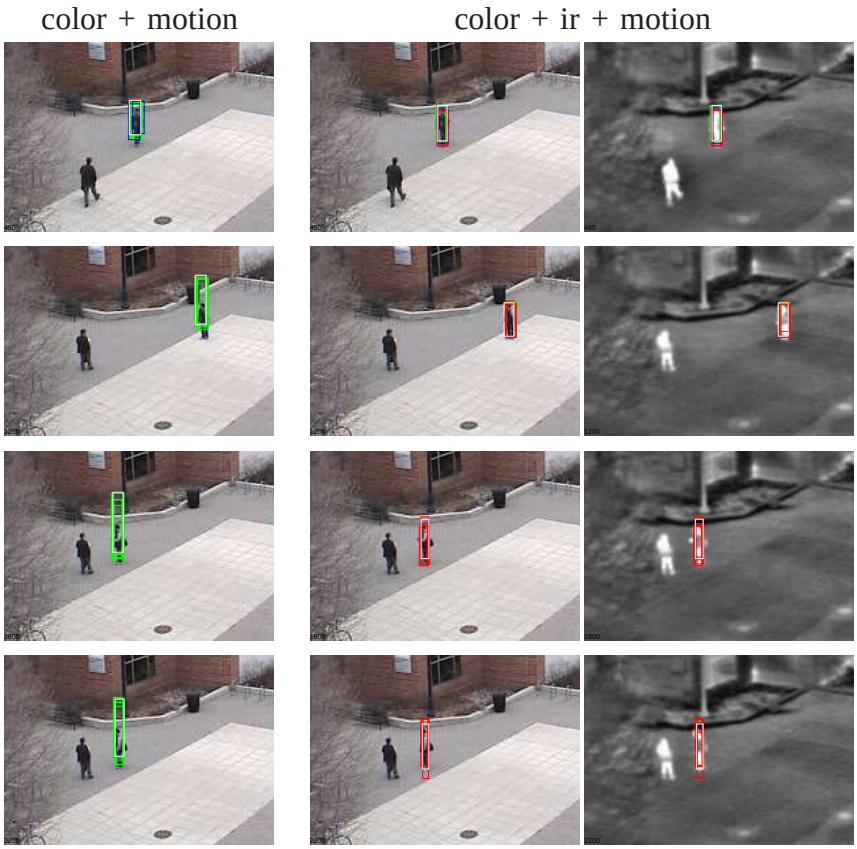


Fig. 13. **seq. #2** Sample tracking results with context-sensitive reliabilities. For the framework built upon color and motion refining the reliabilities with respect to the contextual information does not provide a significant improvement in the outcomes since these cues both fail to account for the uncertainties in the tracking sequence. It does, however, result in more accurate tracking of the person of interest for the framework in which infrared brightness is introduced as a complementary cue as infrared brightness is given a higher importance than the other cues during tracking (see Figure 14).

higher weight than motion during tracking (Figure 19). This experiment shows that combining several visual cues does not always mean robustness. It improves the tracking results only when at least one of the cues considered in tracking is effective in describing the object which is tracked. For instance, in this example, color and motion both fail to account for the uncertainties. It should be added that the two-layered sampling approach suggested in [18] produces much worse tracking results than ours as illustrated in Figure 18 since it relies on first motion and then color information for the localization of the person of interest.

VII. Summary and Future Work

We have presented a particle filter-based tracking algorithm which integrates multiple cues in a novel way. Unlike previous approaches, our method performs the multi-cue integration both in making predictions about the object of interest and in verifying them through observations. Both stages of the integration depend on the reliabilities of the visual cues, which are adapted in a dynamic way. Particularly, in the prediction step, the reliabilities determine to which cue and the proposal function the particles are attached, forcing reliable proposal functions to be employed more in the sequential importance sampling. Moreover, in the measurement step, they specify the level of contribution of each visual cue to the compound likelihood, resulting in more precise weights for the particles.

We have demonstrated the potential of the proposed approach on various illustrative video sequences with different tracking scenarios. As the experimental results reveal, dynamic structure of our formulation makes tracking process easily adapt itself to changes in the context. The proposed framework is general enough to easily include other sources of information. Even though in our experiments we use color, motion

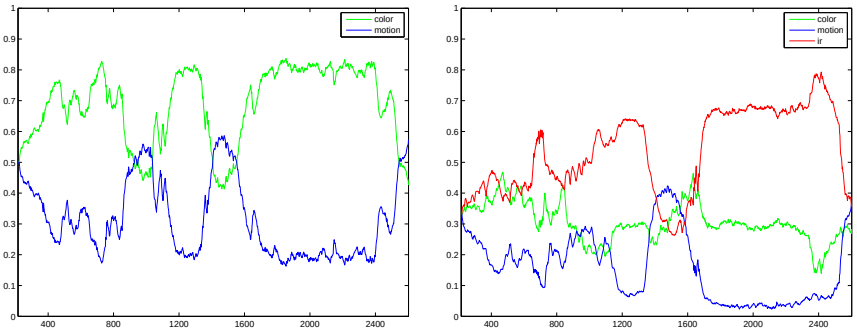


Fig. 14. **seq. #2** Reliabilities throughout the sequence.

Fig. 15. **seq. #3** Sample tracking results based on motion. Employing motion data alone cannot distinguish the actual person of interest from the other persons in motion, and thus results in inaccurate tracking.

Fig. 16. **seq. #3** Sample tracking results based on color. The framework is successful in tracking the person of interest for most part of the sequence, but it loses the track whenever a person having a similar appearance enters the scene.

Fig. 17. **seq. #3** Sample tracking results based on both color and motion with context-sensitive reliabilities. With respect to the contextual information, color is determined to be the main cue for tracking and is given a much higher weight than motion throughout tracking (see Figure 19). Thus, the proposed tracking framework gives outcomes similar to those obtained by using color data alone; it succeeds in tracking the person of interest until a person with a similar appearance appears in the video sequence.

Fig. 18. **seq. #3** Sample tracking results based on the two-layered partitioned sampling approach suggested by Perez et al. [18]. The video sequence involves several persons in motion, and since the localization of the person of interest is depend on primarily motion and then color information, the particles are distributed all around these persons.

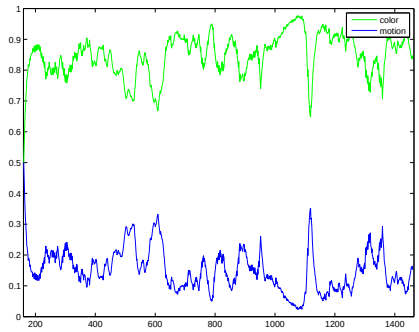


Fig. 19. **seq. #3** Reliabilities throughout the sequence.

and infrared brightness cues as the main sources of information for tracking an object, we can extend this list with further visual cues and integrate them in our framework without any difficulty. Moreover, the suggested approach allows introducing new modalities, whenever available, throughout tracking. However, it is important to note that combining several visual cues does not always increase the tracking accuracy as our last experiment illustrates. Integrating various visual cues does improve the outcomes by eliminating the ambiguities only when at least one of the cues considered in tracking is effective in describing the object of interest, which is not a very surprising result.

In updating the reliabilities of the visual cues, we adopt the approach suggested in [21]. As a future work, it could be interesting to develop new quality measures in updating the cues' reliabilities. For example, one can consider fuzzy measures instead of the hard decision utilized in [21].

REFERENCES

[1] BEHAVE interactions test case scenarios. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS, last accessed: September 2009.

[2] CAVIAR: Context aware vision using image-based active recognition, EC funded CAVIAR project/IST 2001 37540. http://homepages.inf.ed.ac.uk/rbf/CAVIAR, last accessed: September 2009.

[3] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[4] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99109, 1943.

[5] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. Int. Conf. Computer Vision*, pages 232–237, 1998.

[6] P. Brasnett, L. Mihaylova, D. Bull, and N. Canagarajah. Sequential Monte Carlo tracking by fusing multiple cues in video sequences. *Image Vision Comput.*, 25(8):1217–1227, 2007.

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

[8] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Comput. Vis. Image Understand.*, 106(2-3):162–182, 2007.

[9] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. Technical report, Department of Statistics, University of British Columbia, 2008.

[10] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian bayesian state estimation. *IEE Proc. F*, 140(2):107113, 1993.

[11] M. Isard and A. Blake. CONDENSATION-conditional density propagation for visual tracking. *Int. J. Comput. Vis.*, 29(1):5–28, 1998.
[12] M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. Eur. Conf. Computer Vision*, pages 893–908, 1998.
[13] R. A. Jacobs. What determines visual cue reliability? *Trends in Cognitive Sciences*, 6(8):345–350, 2002.
[14] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D, J. Basic Eng.*, 82:35–45, 1960.
[15] E. Maggio, F. Smeraldi, and A. Cavallaro. Combining colour and orientation for adaptive particle filter-based tracking. In *Proc. British Machine Vision Conf.*, pages 659–668, 2005.
[16] K. Nickel and R. Stiefelhagen. Dynamic integration of generalized cues for person tracking. In *Proc. Eur. Conf. Computer Vision*, pages 514–526, 2008.
[17] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proc. Eur. Conf. Computer Vision*, pages 661–675, 2002.
[18] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, 2004.
[19] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Mach. Vision Appl.*, 14(1):50–58, 2003.
[20] J. Triesch, D. H. Ballard, and R. A. Jacobs. Fast temporal dynamics of visual cue integration. *Perception*, 31(4):421–434, 2002.
[21] J. Triesch and C. von der Malsburg. Democratic integration: Self-organized integration of adaptive cues. *Neural Computation*, 13(9):2049–2074, 2001.
[22] Y. Wu and T. S. Huang. A co-inference approach to robust visual tracking. In *Proc. Int. Conf. Computer Vision*, pages 26–33, 2001.

Dépôt légal : 2010 – 1er trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI