Minimum Uncertainty Gap for Robust Visual Tracking

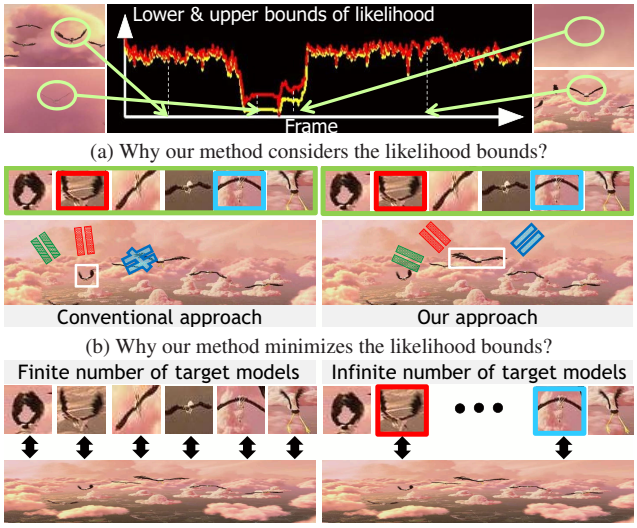
Junseok Kwon and Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University, 151-742, Seoul, Korea
{paradis0, kyoungmu}@snu.ac.kr, http://cv.snu.ac.kr

Abstract

We propose a novel tracking algorithm that robustly tracks the target by finding the state which minimizes uncertainty of the likelihood at current state. The uncertainty of the likelihood is estimated by obtaining the gap between the lower and upper bounds of the likelihood. By minimizing the gap between the two bounds, our method finds the confident and reliable state of the target. In the paper, the state that gives the Minimum Uncertainty Gap (MUG) between likelihood bounds is shown to be more reliable than the state which gives the maximum likelihood only, especially when there are severe illumination changes, occlusions, and pose variations. A rigorous derivation of the lower and upper bounds of the likelihood for the visual tracking problem is provided to address this issue. Additionally, an efficient inference algorithm using Interacting Markov Chain Monte Carlo is presented to find the best state that maximizes the average of the lower and upper bounds of the likelihood and minimizes the gap between two bounds simultaneously. Experimental results demonstrate that our method successfully tracks the target in realistic videos and outperforms conventional tracking methods.

1. Introduction

The objective of the tracking problem is to track the target accurately in the real environments [3, 5, 6, 8, 9, 11, 13, 15, 21, 31, 35, 36]. For robust tracking, most conventional tracking methods formulate the tracking problem by the Bayesian framework [10, 20, 18, 26, 27, 28, 33, 34]. In the Bayesian tracking approach, the goal of the tracking problem is to find the best state, which maximizes the posterior probability $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$. This is called as the Maximum a Posterior (MAP) estimation: $\hat{\mathbf{X}}_t^{MAP} = \arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Y}_{1:t})$, where $\hat{\mathbf{X}}_t^{MAP}$ denotes the best (MAP) state at time t. To obtain the MAP state, the method searches for the state that maximizes the likelihood $p(\mathbf{Y}_t|\mathbf{X}_t)$, which is near the previous state as a prior. In this case, the likelihood is typically calculated by measuring the similarity between the observation $\mathbf{Y}_t$ at the state $\mathbf{X}_t$ and the target model M_t at time t.



(a) Why our method considers the likelihood bounds?

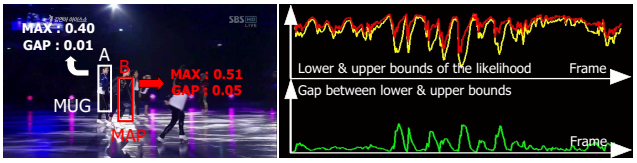
(b) Why our method minimizes the likelihood bounds?

(c) How our method efficiently employs infinite number of target models?

Figure 1. **Basic idea of the proposed method**. (a) The likelihood bounds (uncertainty) are formed inevitably in real tracking situations due to different target models that are employed via different updating strategies during the tracking process. (b) A large gap between the upper and lower bounds indicates that the corresponding state gives very different answers (likelihoods) depending on the target models used (red and blue), although the average likelihood obtained by using set of all target models (green) is high. That means the likelihood estimation over that state is uncertain and unreliable. So, our method tries to find the state that has minimum gap (uncertainty), which gives consistent answers (likelihoods), regardless of the target models. And by maximizing the average likelihood bound at the same time, our method gets the state, which confidently maximizes the likelihood. (c) The proposed method only compares two target models with observations while utilizing the infinite number of target models, which generate lower and upper bounds of the likelihood. On the other hand, other methods compares all the finite number of target models to evaluate the likelihood.

In this case, the MAP estimation assumes that the best state produces the highest likelihood score near by the previous state. However, in the real-world scenario, this assumption is not valid unless the target model M_t is always correct. In practice, the target model is easily corrupted and distorted during the online update. To deal with severe appearance changes, many tracking algorithms evolve the target model with online update. However, because of the

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.305



(a) MAP and MUG (b) Likelihood bounds and their gap

Figure 2. **Example of our tracking results** in *skating1* seq. Our method successfully tracks the target using the MUG estimation, whereas the conventional methods fail to track it using the MAP estimation. The MUG estimation finds the true state A of the target because the gap between the likelihood bounds in State A is smaller than that in State B. On the other hand, the MAP estimation finds the wrong state B because the posterior probability in State B is larger than that in State A.

tracking error, the target model includes more background and becomes erroneous as time goes on. Eventually, the conventional trackers drift into the background and fail to track the target. This drift phenomenon frequently occurs even though the methods find the optimal MAP state because of the noisy target model. According to this fundamental inherent problem, the conventional MAP-based tracking approach need to be reconsidered.

Thus, in this paper, we redefine the goal of tracking problem as to find *the best state that maximizes the average bound of the likelihood and, at the same time, minimizes the gap between bounds of the likelihood*. We call this the *Minimum Uncertainty Gap* (MUG) estimation. Note that in general tracking problem, the upper and lower bounds or the uncertainties of the likelihoods are naturally formed since many different likelihoods are made by different target models that are the reference appearances of the target. The different target models are usually constructed due to the different updating strategies during the tracking process [23]. Specially when there exist severe occlusions, illumination changes, and so on, the likelihood uncertainty becomes larger, as empirically demonstrated in Fig.1(a). This is because the distractors such as occlusions and illumination changes usually make the target models to be much different with each other.

Since the different likelihoods can be generated by different target models, obtaining the likelihood bounds is the same as considering all possible target models that could be constructed. Using the likelihood bounds, the proposed method can find the good target state because it implicitly covers all possible appearance changes of the target with all possible target models. Thus, as illustrated in Fig.1(b), the large gap between the upper and lower bounds indicates that the corresponding state can have either a very good likelihood or a very bad likelihood depending on the employed target model. In this case, the likelihood estimation over the state is easily affected by the noisy target models and the estimated likelihood is uncertain and not reliable. Hence, by minimizing the gap between the two likelihood bounds, the proposed method can find the confident state of the target. MUG is also affected by aforementioned distractors (outliers) in the target model. Nevertheless, MUG is more robust to them. This is because MUG provides the confidence score about the likelihood estimation, whereas MAP cannot. Note that the outliers usually produce low confidence scores [7]. Therefore, MUG can easily identify and avoid them by estimating the confidence values of them. To measure the confidence of the likelihood, our method estimates the lower and upper bounds of the likelihood, minimizes the gap between the bounds, and accurately tracks the target, as shown in Fig.2.

In tracking methods using multiple target models, the VTS tracker [19] tracked the target successfully with the visual tracker sampler framework. The MIL tracker [2] solved the ambiguity of the target appearance using multiple instances of the target model and robustly tracked the target with the online multiple instance learning algorithm. Compared with these works which utilized a relatively small number of target models, our method implicitly employs all possible target models with the likelihood bounds. In tracking methods with the likelihood bound, the L1BPR tracker [24] proposed an efficient L1 tracker with the Bounded Particle Re-sampling (BPR) technique which considers the upper bound of the likelihood. However, the method used the BPR technique to speed up the L1 tracker without sacrificing accuracy. Our method utilizes the likelihood bounds to measure uncertainty of the likelihood and enhances the accuracy of visual tracking. In sampling based tracking methods, the particle filter [12] handled the non-Gaussianity of the target distribution in the tracking problems. In [16, 32], the Markov Chain Monte Carlo (MCMC) method coped with the high-dimensional state spaces, whereas the joint particle filter was not able to. The Interacting Markov Chain Monte Carlo (IMCMC) method [19] required a relatively small number of samples by exchanging information about good states between Markov Chains. Our method employs the IMCMC method. However, our method uses the samples to obtain the state that minimizes the uncertainty of the target distribution, whereas the samples in the other methods are used to obtain the maximum of the target distribution.

The first contribution of the paper is a novel tracking framework that utilizes the *MUG* instead of the MAP estimation. By finding the state that minimizes the gap between the likelihood bounds, our method can cope with the drift problem caused by a noisy target model more robustly than the conventional MAP-based approach, and successfully tracks the target when there are severe illumination changes, occlusions, and pose variations. The second contribution is a rigorous derivation of the lower and upper bounds of the likelihood in the visual tracking problem. Although the bounds of the likelihood are obtained based on [14], applying those bounds into the visual tracking problem directly is not trivial since proper and careful

designs of the parameter γ and the distribution q are needed for the visual tracking problem. In this work, γ and q are designed as the hyper parameter of the likelihood and the prior distribution of a target model, respectively. The last contribution is an efficient strategy to obtain the state that has the Minimum Uncertainty Gap. Our method constructs two chains and inferences the best state on the chains using the IMCMC method in [19]. In the first chain, the proposed method finds the state that maximizes the average bound (mean of the lower and upper bounds) of the likelihood. In the second chain, the method searches for the state that minimizes the gap between two likelihood bounds. These chains communicate with each other to obtain the best state that maximizes the average bound and minimizes the gap between bounds at the same time.

2. New Objective of the Bayesian Tracker

To find the best state, $\hat{\mathbf{X}}_t$, our method obtains the Minimum Uncertainty Gap (MUG) at each time t as follows:

$$\hat{\mathbf{X}}_t = \arg\min_{\mathbf{X}_t} \frac{p_u(\mathbf{Y}_t|\mathbf{X}_t) - p_l(\mathbf{Y}_t|\mathbf{X}_t)}{p_u(\mathbf{Y}_t|\mathbf{X}_t) + p_l(\mathbf{Y}_t|\mathbf{X}_t)}, \quad (1)$$

where $p_l(\mathbf{Y}_t|\mathbf{X}_t)$ and $p_u(\mathbf{Y}_t|\mathbf{X}_t)$ denote the lower and upper bounds of the likelihood, respectively. In (1), our method finds the state that maximizes the average bound $[p_u(\mathbf{Y}_t|\mathbf{X}_t) + p_l(\mathbf{Y}_t|\mathbf{X}_t)]$ and minimizes the gap between bounds, $[p_u(\mathbf{Y}_t|\mathbf{X}_t) - p_l(\mathbf{Y}_t|\mathbf{X}_t)]$, at the same time. The best state (MUG state) at time t is represented as a three-dimensional vector $\hat{\mathbf{X}}_t = (\hat{X}_t^x, \hat{X}_t^y, \hat{X}_t^s)$, where $\hat{X}_t^x, \hat{X}_t^y$, and $\hat{X}_t^s$ indicate the x, y position and the scale of the target, respectively.

To obtain the MUG state, we need to estimate the lower and upper bounds of the likelihood. First, we define the likelihood as

$$p(\mathbf{Y}_t, \theta|\mathbf{X}_t) = exp^{-\lambda dist\left(\theta, \mathbf{Y}_t(\mathbf{X}_t)\right)}, \quad (2)$$

where $\mathbf{Y}_t(\mathbf{X}_t)$ denotes the observation at the state $\mathbf{X}_t$, $dist\big(\theta, \mathbf{Y}_t(\mathbf{X}_t)\big)$ represents the dissimilarity measure between the target model θ and the observation $\mathbf{Y}_t(\mathbf{X}_t)$, and λ is a weighting parameter. The observation and the target model are modeled by HSV histogram. The dissimilarity measure is designed by Bhattacharyya similarity coefficient [28]. As aforementioned, the main cause of the tracking failures is the noisy target models. Therefore, our method integrates out the target model θ in (2) and estimates the log marginal likelihood: $\int_\Theta \ln p(\mathbf{Y}_t, \theta|\mathbf{X}_t)d\theta$, where Θ denotes the whole target model space. To approximate the integral numerically, we obtain the mathematical lower (Jensen's inequality) and mathematical upper bounds (Gibbs' inequality) of the marginal likelihood based on [14].

$$\ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma) = \int_\Theta q(\theta|\gamma, \mathbf{X}_t) \ln \frac{p(\mathbf{Y}_t, \theta|\mathbf{X}_t)}{q(\theta|\gamma, \mathbf{X}_t)} d\theta, \quad (3)$$

$$\ln p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma) = \int_\Theta p(\theta|\mathbf{Y}_t, \mathbf{X}_t) \ln \frac{p(\mathbf{Y}_t, \theta|\mathbf{X}_t)}{q(\theta|\gamma, \mathbf{X}_t)} d\theta, \quad (4)$$

where $q(\theta|\gamma, \mathbf{X}_t)$ is the prior distribution of the target model θ and γ is the hyper parameter of the distribution. Because θ is marginalized out in (3)(4), the lower and upper bounds of the likelihood is the function of $\mathbf{X}_t$ and γ, which are a state and a parameter, respectively. Then, the goal of our method is to find both the best state and parameter, which reduce gap between the likelihood bounds. Thus, our method composes two main parts as follows:

• **Parameter learning (Section 3):** Using the MUG states, $\{\hat{\mathbf{X}}_i\}_{i=1}^t$, our method learns the parameter γ for time $t+1$. In our method, the parameter is not set empirically but is obtained analytically to maximize the lower bound in (3) and to minimize the upper bound in (4). Moreover, the parameter is not fixed to constant but is adaptively varied at each time t by the process in Section 3.

• **State inference (Section 4):** Given the parameter γ estimated at time $t-1$, our method finds the MUG state $\hat{\mathbf{X}}_t$ at time t, which produces the minimum uncertainty gap. To achieve the goal, the method searches states that maximize the average bound by increasing the denominator in (1). Thus, the method can obtain good quality of states with high likelihood scores. This advantage is similar to that of the MAP estimation. In addition, it prevents the best state with the minimum uncertainty gap from having a low likelihood score. Our method simultaneously searches states that minimize uncertainty of the likelihood estimation by decreasing the numerator in (1). Then, the method can avoid outliers which have a large uncertainty gap, even though they have high likelihood scores. This advantage cannot be achieved in the MAP estimation.

3. Parameter Learning

3.1. Learning γ_u for the Upper Bound

We learn the best parameter γ_u which minimizes the upper bound (4): $\gamma_u = \arg\min_\gamma \ln p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma)$. Then, γ_u also minimizes the KL divergence $D(p \parallel q)$ because $\ln p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma) = D(p \parallel q) + \ln p(\mathbf{Y}_t|\mathbf{X}_t)$, where

$$D(p \parallel q) = \int_\Theta p(\theta|\mathbf{Y}_t, \mathbf{X}_t) \ln \frac{p(\theta|\mathbf{Y}_t, \mathbf{X}_t)}{q(\theta|\gamma, \mathbf{X}_t)} d\theta. \quad (5)$$

The parameter γ_u that minimizes $D(p \parallel q)$ satisfies $\mathbb{E}_{q(\theta|\gamma_u, \mathbf{X}_t)}[v(\theta)] = \mathbb{E}_{p(\theta|\mathbf{Y}_t, \mathbf{X}_t)}[v(\theta)]$, as derived in Appendix (supplementary material). This means that the minimization of KL divergence is equivalent to Moment Matching (MM) of θ [4]: the first and second moments of θ under the distribution $q(\theta|\gamma_u, \mathbf{X}_t)$ is equal to those under the distribution $p(\theta|\mathbf{Y}_t, \mathbf{X}_t)$. In (5), the prior $q(\theta|\gamma, \mathbf{X}_t)$ is de-

signed as

$$q(\theta|\gamma, \mathbf{X}_t) = q(\theta|\mu, \sigma, \mathbf{X}_t) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\left(\frac{\theta-\mu}{\sqrt{2}\sigma}\right)^2}, \quad (6)$$

so the parameter $\gamma_u = (\mu_u, \sigma_u)$ is obtained for each bin in the HSV histogram:

• **1st MM:** Since the first moment of θ under $q(\theta|\gamma_u, \mathbf{X}_t)$ and under $p(\theta|\mathbf{Y}_t, \mathbf{X}_t) \approx p(\mathbf{Y}_t, \theta|\mathbf{X}_t)$ in (2) is μ_u and $\int_\Theta \theta p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta$, respectively, the following can be taken:

$$\mu_u = \int_\Theta \theta p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta. \quad (7)$$

In (7), the integration over θ is approximated using the Z samples of θ, $\{\theta_i\}_{i=1}^Z$, where θ_i is designed as $\mathbf{Y}_i(\hat{\mathbf{X}}_i)$, which indicates the observation around the MUG state at the *i*-th recent frame. By substituting $\mathbf{X}_t = \hat{\mathbf{X}}_t$ and $p(\theta|\mathbf{Y}_t, \mathbf{X}_t) \approx p(\mathbf{Y}_t, \theta|\mathbf{X}_t)$ in (2) into (7), we get $\mu_u = \int_\Theta \theta p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta \approx \frac{1}{Z}\sum_{i=t-Z}^{t-1} \theta_i exp^{-\lambda dist\left(\theta_i, \mathbf{Y}_t(\hat{\mathbf{X}}_t)\right)}$.

• **2nd MM:** Since the second moment of θ under $q(\theta|\gamma_u, \mathbf{X}_t)$ and $p(\theta|\mathbf{Y}_t, \mathbf{X}_t)$ is σ_u and $\int_\Theta \left(\theta - \mu_u\right)^2 p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta$, respectively, the following can be taken:

$$\sigma_u = \int_\Theta \left(\theta - \mu_u\right)^2 p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta, \quad (8)$$

where $\int_\Theta \left(\theta - \mu_u\right)^2 p(\theta|\mathbf{Y}_t, \mathbf{X}_t)d\theta \approx \frac{1}{Z}\sum_{i=t-Z}^{t-1} \left(\theta_i - \mu_u\right)^2 exp^{-\lambda dist\left(\theta_i, \mathbf{Y}_t(\hat{\mathbf{X}}_t)\right)}$.

Finally, the global minimum of the upper bound of the likelihood at the state $\mathbf{X}_t$ in (4) is estimated based on [14]:

$$\ln p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) \approx \frac{1}{Z}\sum_{i=t-Z}^{t-1} \ln \frac{p(\theta_i, \mathbf{Y}_t|\mathbf{X}_t)}{q(\theta_i|\gamma_u, \mathbf{X}_t)}, \quad (9)$$

where the integration in (4) is approximately obtained using the Z samples of θ, $\{\theta_i\}_{i=1}^Z$, where θ_i indicates the observation around the MUG state at the *i*-th recent frame.

3.2. Learning γ_l for the Lower Bound

We learn the best parameter γ_l which maximizes the lower bound in (3): $\gamma_l = \underset{\gamma}{\operatorname{argmax}} \ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma)$. For this purpose, the gradient of $\ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma)$ is taken with respect to γ to zero:

$$\frac{d}{d\gamma} \ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma) = -\int_\Theta h(\theta|\gamma, \mathbf{X}_t) q(\theta|\gamma, \mathbf{X}_t)d\theta = 0, \quad (10)$$

where

$$h(\theta|\gamma, \mathbf{X}_t) = h(\theta|\mu, \sigma, \mathbf{X}_t) = \left[1 + \ln \frac{q(\theta|\mu, \sigma, \mathbf{X}_t)}{p(\mathbf{Y}_t, \theta|\mathbf{X}_t)}\right] \left[\begin{array}{c} \frac{\partial}{\partial\mu} q(\theta|\mu, \sigma, \mathbf{X}_t) \\ \frac{\partial}{\partial\sigma} q(\theta|\mu, \sigma, \mathbf{X}_t) \end{array}\right]. \quad (11)$$

[1] To find the parameter $\gamma_l = (\mu_l, \sigma_l)$ that satisfies (10), the quasi-optimized lower bound is estimated by Stochastic Approximation Monte Carlo (SAMC) in [22] to define the recursive approximation of the solution of $\frac{d}{d\gamma} \ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma) = 0$. SAMC then iteratively updates a sequence of values via the recursion:

$$\begin{aligned} \gamma^{(n+1)} &= \gamma^{(n)} + s^{(n+1)}\omega(\gamma^{(n)}), \\ \omega(\gamma^{(n)}) &= -\int_\Theta h(\theta|\gamma^{(n)}, \mathbf{X}_t)d\theta, \end{aligned} \quad (12)$$

where $\int_\Theta h(\theta|\gamma^{(n)}, \mathbf{X}_t)d\theta \approx \frac{1}{Z}\sum_{i=t-Z}^{t-1} h(\theta_i|\gamma^{(n)}, \hat{\mathbf{X}}_t)$. $\gamma^{(n+1)}$ denotes approximation of the γ value at *(n+1)*-th iteration, and $s^{(n+1)}$ indicates the modification factor at *(n+1)*-th iteration, which linearly decreases from 0.5 to 0.1 as time goes on. After the predefined iterations N, we get $\gamma^{(N)} = \gamma_l = (\mu_l, \sigma_l)$.

Then, the final estimate of the lower bound at the state $\mathbf{X}_t$ in (3) is estimated based on [14]:

$$\ln p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l) \approx \frac{1}{Z}\sum_{i=t-Z}^{t-1} \ln \frac{p(\theta_i, \mathbf{Y}_t|\mathbf{X}_t)}{q(\theta_i|\gamma_l, \mathbf{X}_t)}. \quad (13)$$

4. Inference of the MUG State

To find the best state that satisfies (1) with the fixed parameters γ_l and γ_u, our method utilizes the IMCMC sampling method [19]. In the IMCMC sampling method, two markov chains are designed. The first chain frequently accepts the state that maximizes the average likelihood bound. The second frequently accepts the state that minimizes the gap between the bounds. The IMCMC sampling method consists of two modes, parallel and interacting. In the parallel mode, our method acts as the parallel Metropolis Hastings algorithm and separately obtains samples over those chains via two main steps: the proposal step and the acceptance step. At the proposal step, a new state is proposed by the proposal density function.

$$Q_(\mathbf{X}_t^*; \mathbf{X}_t) = G(\mathbf{X}_t, \sigma_p^2), \quad (14)$$

where Q denotes the proposal density function, G represents the Gaussian distribution with mean $\mathbf{X}_t$ and variance σ_p^2, and $\mathbf{X}_t^*$ represents a new state at time t. Given the proposed state, each chain decides whether the state is accepted or not with the acceptance ratio in the acceptance step:

$$\begin{aligned} a_1^p &= min\left[1, \frac{\left[p_u(\mathbf{Y}_t|\mathbf{X}_t^*, \gamma_u) + p_l(\mathbf{Y}_t|\mathbf{X}_t^*, \gamma_l)\right] Q(\mathbf{X}_t; \mathbf{X}_t^*)}{\left[p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) + p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)\right] Q(\mathbf{X}_t^*; \mathbf{X}_t)}\right], \\ a_2^p &= min\left[1, \frac{\left[p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) - p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)\right] Q(\mathbf{X}_t; \mathbf{X}_t^*)}{\left[p_u(\mathbf{Y}_t|\mathbf{X}_t^*, \gamma_u) - p_l(\mathbf{Y}_t|\mathbf{X}_t^*, \gamma_l)\right] Q(\mathbf{X}_t^*; \mathbf{X}_t)}\right], \end{aligned} \quad (15)$$

[1] In (11), $[1 + \ln \frac{q(\theta|\mu,\sigma,\mathbf{X}_t)}{p(\mathbf{Y}_t,\theta|\mathbf{X}_t)}] = [\lambda dist\left(\theta, \mathbf{Y}_t(\mathbf{X}_t)\right) - \left(\frac{\theta-\mu}{\sqrt{2}\sigma}\right)^2 - \ln\sqrt{2\pi\sigma^2} + 1]$, $\frac{\partial}{\partial\mu} q(\theta|\mu, \sigma, \mathbf{X}_t) = \frac{1}{\sqrt{2\pi}\sigma} exp^{-\left(\frac{\theta-\mu}{\sqrt{2}\sigma}\right)^2} \frac{1}{\sigma^2}(\theta - \mu)$, and $\frac{\partial}{\partial\sigma} q(\theta|\mu, \sigma, \mathbf{X}_t) = -\frac{1}{2\sigma^2} + \frac{(\theta-\mu)^2}{2\sigma^2}$.

Algorithm 1 Minimum Uncertainty Gap tracker

Input: $\mathbf{X}_{t-1} = (\mathbf{X}^x_{t-1}, \mathbf{X}^y_{t-1}, \mathbf{X}^s_{t-1}), \alpha = 1$

Output: $\hat{\mathbf{X}}_t = (\hat{\mathbf{X}}^x_t, \hat{\mathbf{X}}^y_t, \hat{\mathbf{X}}^s_t)$

while all frames **do**
  for 1 to R **do**
    Choose mode. Sample $\rho \sim U[0, 1]$.
    if $\rho < \alpha$ **then**
      Interacting mode:
      Two chains replace their states with that of the other with the probability a^i_1 and a^i_2 in (16), respectively.
    else
      Parallel mode:
      Two chains propose states using Q in (14) and accepts the states with the probability a^p_1 and a^p_2 in (15), respectively.
    end if
    Decrease the α value.
  end for
  Estimate the MUG state, $\hat{\mathbf{X}}_t$ using (1).
  Determine γ_u using (7)(8) and γ_l using (12).
end while

where $p_l(\mathbf{Y}_t|\mathbf{X}^*_t, \gamma_l)$ and $p_u(\mathbf{Y}_t|\mathbf{X}^*_t, \gamma_u)$ denote the estimated lower and upper bounds of the likelihood at the state $\mathbf{X}^*_t$, respectively. These steps iteratively proceed until the number of iterations reaches the predefined value R.

When the method is in the interacting mode, the trackers communicate with the others and make leaps to better states of the target. Due to the interaction mode, our method can find the common state, which maximizes the average likelihood bound and, at the same time, minimizes the gap between bounds. A chain accepts the state of the chain 1 as its own state with the probability a^i_1, if the state of the chain 1 greatly maximizes the average likelihood bound. Similarly, a chain accepts the state of the chain 2 as its own state with the probability a^i_2, if the state of the chain 2 greatly minimizes the gap between bounds:

$$
\begin{aligned}
a^i_1 &= \frac{[p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) + p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)] - \Delta_1}{2p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l) - \Delta_1 + \Delta_2}, \\
a^i_2 &= \frac{\Delta_2 - [p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) - p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)]}{2p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l) - \Delta_1 + \Delta_2}, \\
\Delta_1 &= \text{MAX}\left(\left[p_u(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1}) + p_l(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1})\right] - \frac{1}{4}, 0\right), \\
\Delta_2 &= \text{MIN}\left(\left[p_u(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1}) - p_l(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1})\right] + \frac{1}{4}, 1\right).
\end{aligned}
\tag{16}
$$

In (16), $[p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) + p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)] - \Delta_1$ indicates the increased quantity of the average likelihood bound and $\Delta_2 - [p_u(\mathbf{Y}_t|\mathbf{X}_t, \gamma_u) - p_l(\mathbf{Y}_t|\mathbf{X}_t, \gamma_l)]$ represents the decreased quantity of the gap between bounds. $p_l(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1})$ and $p_u(\mathbf{Y}_t|\hat{\mathbf{X}}_{t-1})$ are the lower and upper bounds of the likelihood on the best state at the previous frame with the best parameter, respectively. Our method operates in the interacting mode with the probability α, which linearly decreases from 1.0 to 0.0 as the simulation goes on. Notably, the IMCMC method [19] typically converges to the invariant distribution $\frac{p_u(\mathbf{Y}_t|\mathbf{X}_t) - p_l(\mathbf{Y}_t|\mathbf{X}_t)}{p_u(\mathbf{Y}_t|\mathbf{X}_t) + p_l(\mathbf{Y}_t|\mathbf{X}_t)}$ in (1). Algorithm 1 illustrates whole process of our method.

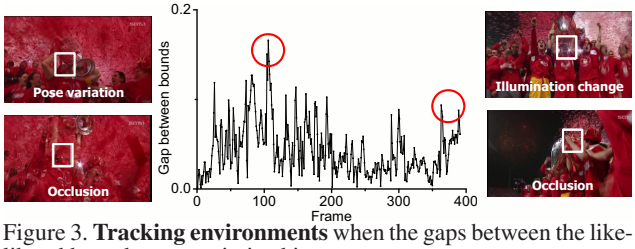
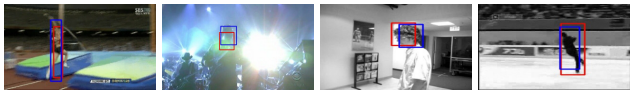


Figure 3. **Tracking environments** when the gaps between the likelihood bounds are maximized in *soccer* seq.



(a) *high-jump* (b) *shaking* (c) *david* (d) *skater*

Figure 4. **States** of the target, which produce the maximum lower bound (blue rectangle) and the minimum upper bound (red rectangle) of the likelihood at a frame.

5. Experimental Results

For the experiments, publicly available video sequences obtained from [1, 17, 19, 25, 34, 30] were utilized. Using the sequences, the proposed method (MUG) was analyzed and compared with 6 state-of-the-art tracking methods, MC [16, 28], IVT [29], FRAGT [1], MIL [2], VTS [19], and MTT [37]. In all experiments, λ in (2) is set to 5. Z in (9) and (13) is set to 15. σ_p in (14) is set to $\sigma^x_p = \sqrt{4}, \sigma^y_p = \sqrt{2}$, and $\sigma^s_p = \sqrt{0.01}$, where σ^x_p, σ^y_p, and σ^s_p denote the variance of the x translation, y translation and the scale, respectively. Please note that **our method always use the same settings throughout all experiments** and the parameters of other methods were adjusted to show the best tracking performance. Same initializations were set to all methods for fair comparison. The software provided by the authors were used to obtain the tracking results of IVT, MIL, FRAGT, VTS, and MTT. The supplementary material contains result videos.

5.1. Analysis of the Proposed Method

Lower and Upper Bounds of the Likelihood: The tracking environments are examined when the gaps between the likelihood bounds are maximized. As illustrated in Fig.3, the gaps between the likelihood bounds were maximized when there were severe occlusions, pose variations, or illumination changes. These changes caused the target appearance to become noisy. Because of the noisy appearance, the estimated likelihood became very uncertain. This uncertainty of the likelihood produced the large gap between the lower and upper bounds of the likelihood in our method. In other words, the likelihood cannot be uniquely determined, especially when the tracking environments include the aforementioned appearance changes. Therefore, our method considers the uncertainty of the likelihood to track the target robustly in real-world situations.

The states of the target that produce the maximum lower bound or the minimum upper bound of the likelihood are

Table 1. **Comparison of tracking results** using MAP. ML, and MUG. The numbers indicate the average center location errors in pixels. The improvement score is calculated by dividing the tracking error of ML3 by that of MUG.

	bird1	*bird2*	*lemming*	*woman*	*soccer*	*skating1*	*diving*	*high-jump*	*skater*
MAP	199	45	11	127	51	115	26	70	30
ML1	208	47	12	137	56	110	43	65	47
ML2	201	51	16	138	61	107	46	71	51
ML3	210	42	11	101	49	150	27	71	37
MUG	13	11	16	14	32	17	14	30	17
Score	**16**	**3.8**	**0.7**	**7.2**	**1.5**	**8.8**	**1.9**	**2.3**	**2.2**

Table 2. **Comparison of tracking results** using MCMC and IMCMC. The improvement score is calculated by dividing the tracking error of MCMC by that of IMCMC.

	bird1	*bird2*	*lemming*	*woman*	*soccer*	*skating1*	*diving*	*high-jump*	*skater*
MCMC	24	31	52	40	47	89	32	65	51
IMCMC	13	11	16	14	32	17	14	30	17
Score	**1.8**	**2.8**	**3.3**	**2.9**	**1.5**	**5.2**	**2.3**	**2.2**	**3.0**

also checked. In (7)(8), the target model for the minimum upper bound is constructed by averaging the target appearance in the recent frames. So, the model is adequate to track the target whose appearance smoothly changes over time. Then, the state that produces the minimum upper bound is the best state when the smooth changes in the target appearance are assumed, as shown in Fig.4(a)(c). In (10)(11), the target model for the maximum lower bound is heavily updated if the current observation is vastly different from the model. So, the model is robust to track the target whose appearance abruptly changes at a certain time. Then, the state that produces the maximum lower bound is the best state when the abrupt changes in the target appearance are assumed, as shown in Fig.4(b)(d). Therefore, the state that reduces the gap between two bounds is the best state for both smooth and abrupt changes in target appearance.

Performance of MUG and IMCMC: To evaluate the performance of MUG, we used the same likelihood function that employs the Bhattacharyya coefficient as the similarity measure and the HSV color histogram as the feature. The only difference is how to determine the best state. The best states estimated by MAP is the one which maximize the posterior probability. The best states estimated by ML1, ML2, and ML3 are the ones which maximize the lower likelihood bound, upper likelihood bound, and average likelihood bound, respectively. The best state obtained by MUG is the one which maximizes the average likelihood bound and simultaneously minimizes the gaps between two bounds. As shown in Table 1, the best state obtained by MUG gives most accurate tracking results. These results demonstrated the state which produces the maximum likelihood score and the maximum posterior probability do not always correspond to the true target state in real-world settings. Additionally, the results shows the methods should consider uncertainty of the estimated likelihood by measuring the gaps between the likelihood bounds like our method.

To evaluate the performance of IMCMC, the same MUG strategy is used to determine the best state. The only difference is the procedure in finding the best state. The best state in MCMC is found by using a single chain, in which the chain finds the state that maximizes the average bound and minimizes the gap between two bounds simultaneously. The best state in IMCMC is obtained by employing two chains, in which one chain only searches for the state that maximizes the average bound and the other only searches for the state that minimizes the gap between two bounds. Then, two chains exchange information about good states. As indicated in Table 2, using two chains shows better tracking performance because the tracking methods using a single chain get trapped in local optima more frequently as the target distribution becomes complex. The target distribution is complex because the different two types of the likelihood distribution are mixed in a single distribution. Our method divides a complex distribution into two simple ones using IMCMC, where two distributions describe the average bound and the gap between two bounds, respectively.

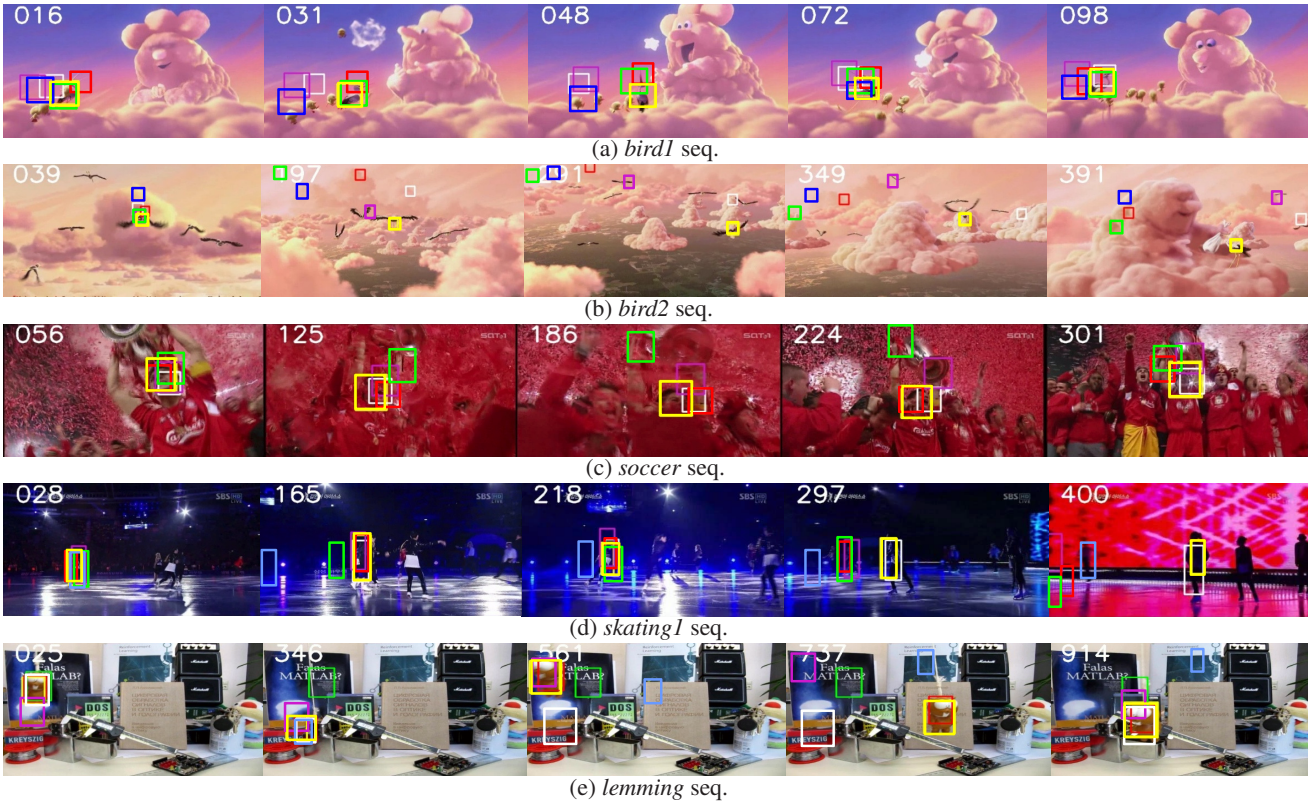
Table 3. **Comparison of tracking results.** The numbers indicate the average center location errors in pixels. Red is the best result and blue is the second-best. Other numbers in () indicate the percent of successfully tracked frames, where tracking is success when the overlap ratio between the predicted bounding box A_p and ground truth bounding box A_g is over than 0.5: $\frac{area(A_p \bigcap A_g)}{area(A_p \bigcup A_g)} > 0.5$.

	MC	**IVT**	**FRAGT**	**MIL**	**VTS**	**MTT**	**MUG**
bird1	215 (16)	230 (13)	228 (13)	270 (11)	**119** (13)	265 (12)	**13** (43)
bird2	40 (18)	115 (11)	24 (67)	**13** (81)	81 (23)	76 (21)	**11** (86)
lemming	**12** (85)	**14** (79)	84 (26)	**14** (51)	70 (45)	90 (25)	16 (71)
woman	138 (11)	133 (11)	112 (19)	120 (15)	**111** (19)	120 (15)	**14** (62)
soccer	53 (15)	116 (9)	82 (11)	41 (17)	**15** (35)	**17**(34)	32 (20)
skating1	172 (14)	213 (11)	93 (26)	85(31)	**8**(93)	150 (20)	**17**(40)
diving	**27** (24)	79 (20)	64 (20)	73 (20)	80 (20)	98 (19)	**14** (24)
high-jump	73 (15)	79 (15)	**69** (15)	91 (15)	143 (14)	94 (15)	**30** (17)
skater	28 (47)	86 (41)	**23** (61)	85 (41)	25 (66)	25 (66)	**17** (66)
Speed(fps)	**7**	**6**	**2**	**17**	**0.4**	**0.1**	**3**

5.2. Comparison with other Tracking methods

As summarized in Table 3, our method (MUG) most accurately tracked the targets in terms of the center location error and the success rate, even though there are several types of appearance changes. VTS showed the second-best tracking performance. Our method was robust to the geometric appearance changes of the non-rigid target in *diving*, *high-jump*, and *skater*; the occlusions in *bird1*, and *woman*; and the motion blur in *bird2*. In this paper, we wanted to demonstrate that our method can produce better tracking results by utilizing a very simple likelihood function and its lower and upper bounds. In using the simple likelihood function, the method was much faster and more accurate than VTS. The tracking performance of our method can be further enhanced if more advanced likelihood functions are employed. Additionally, with the simple likelihood function, our method produced more accurate tracking results than other state-of-the-art methods, where they are robust to the pose variations, occlusions, and illumination changes. Note that, for the sampling-based methods, we used the same number of samples to track the target.

Fig.5 and Fig.6 demonstrate how our method outper-



(a) *bird1* seq.

(b) *bird2* seq.

(c) *soccer* seq.

(d) *skating1* seq.

(e) *lemming* seq.

Figure 5. **Tracking results** in several challenging sequences. Yellow, blue, white, purple, green and red rectangles represent tracking results of MUG, MTT, VTS, MIL, FRAGT, and MC, respectively.

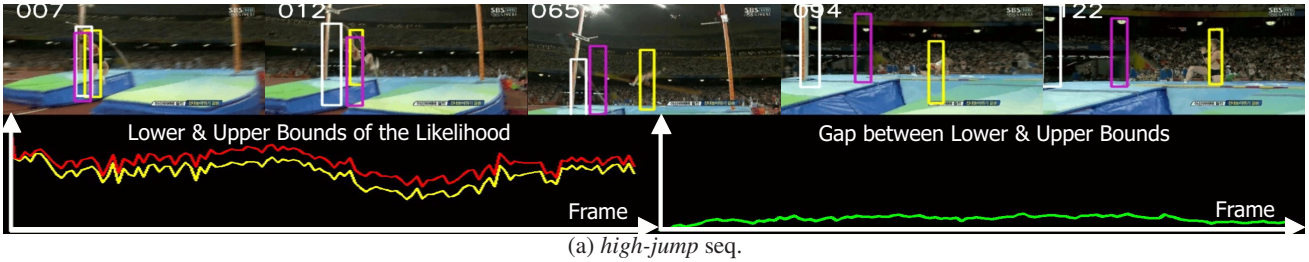
forms the state-of-the-art methods in several challenging sequences. In Fig.5, the tracking performance under the severe occlusions and background clutters was tested. When the sequence contained several appearance changes of the target at the same time, our method robustly tracked the target over time, while other tracking methods frequently missed the targets. The tracking results of MIL drifted into the background when the aforementioned changes transformed the target appearance into a different one. Our method overcame this problem and successfully tracked the target by evaluating the target configuration with several likelihoods. In Fig.6, our method did not miss the target in all frames, although the sequences include the severe geometric appearance changes of the target. On the other hand, MIL and VTS frequently failed to track the target when the target was severely deformed. Our method was more efficient than VTS in terms of the computational cost because it utilized two likelihoods only for evaluating the configuration by estimating the lower and upper bounds of the likelihood.

6. Conclusion

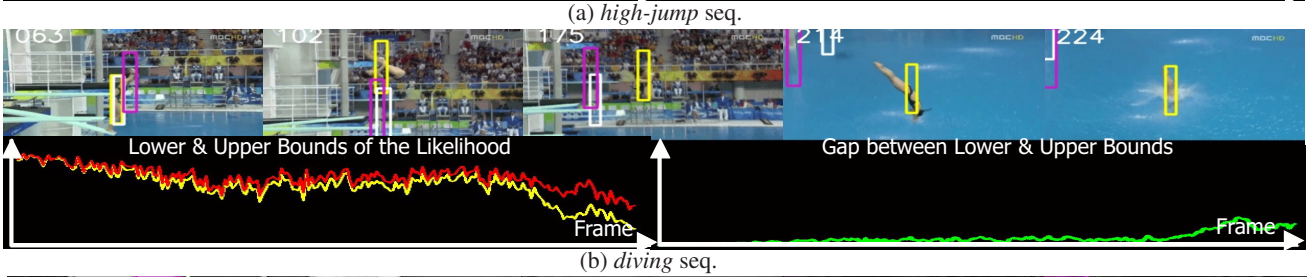
In this paper, we propose a novel tracking framework that tracks the target robustly by finding the best state of the target, which minimizes the gap between the lower and upper bounds of the likelihood. Obtaining the likelihood bounds is the same as considering all possible target models during the tracking process. Therefore, our method finds the good state of the target by reflecting all possible appearance changes of the target. The experimental results demonstrate that our method outperforms the conventional tracking methods using the MAP and ML estimation. The method also shows better tracking performances than those of the state-of-the-art tracking methods when there are illumination, occlusions, and deformation.

References

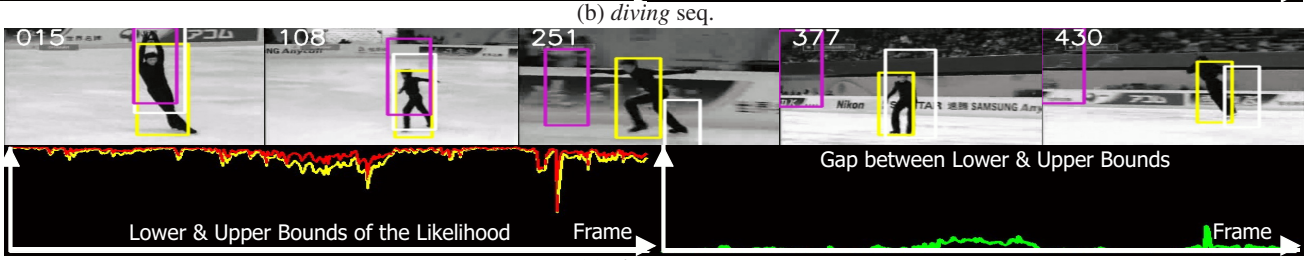
[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. *CVPR*, 2006.
[2] B. Babenko, M. Yang, and S. Belongie. Visual tracking with online multiple instance learning. *CVPR*, 2009.
[3] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. *CVPR*, 1998.
[4] C. I. Byrnes and A. Lindquist. A convex optimization approach to generalized moment problems. *Control and Modeling of Complex Systems, Springer*, 2003.
[5] R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *PAMI*, 27(10):1631–1643, 2005.
[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *CVPR*, 2000.
[7] A. C. Courville, N. D. Daw, G. J. Gordon, and D. S. Touretzky. Model uncertainty in classical conditioning. *NIPS*, 2003.
[8] M. Godec, P. M. Roth, , and H. Bischof. Hough-based tracking of non-rigid objects. *ICCV*, 2011.
[9] M. G. H. Grabner and H. Bischof. Real-time tracking via on-line boosting. *BMVC*, 2006.
[10] B. Han and L. Davis. On-line density-based appearance modeling for object tracking. *ICCV*, 2005.
[11] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. *ICCV*, 2011.



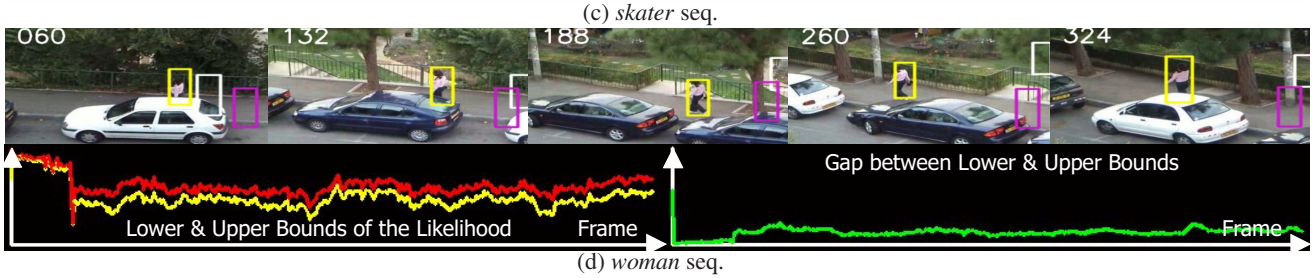
(a) *high-jump* seq.



(b) *diving* seq.



(c) *skater* seq.



(d) *woman* seq.

Figure 6. **Tracking results** with lower and upper bounds of the likelihood obtained by MUG. Yellow, white, and purple rectangles represent tracking results of MUG, VTS, and MIL, respectively. Yellow and red curves represent lower and upper bounds of the likelihood over time in MUG, respectively. Green curve represents gap between the bounds over time in MUG.

[12] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. *ECCV*, 1998.

[13] A. D. Jepson, D. J. Fleet, and T. F. E. Maraghi. Robust online appearance models for visual tracking. *PAMI*, 25(10):1296–1311, 2003.

[14] C. Jia, H. Shenb, and M. Westc. Bounded approximations for marginal likelihoods. *Technical report*, 2010.

[15] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *PAMI*, 34(7):1409–1422, 2012.

[16] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *PAMI*, 27(11):1805–1918, 2005.

[17] J. Kwon and K. M. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. *CVPR*, 2009.

[18] J. Kwon and K. M. Lee. Visual tracking decomposition. *CVPR*, 2010.

[19] J. Kwon and K. M. Lee. Tracking by sampling trackers. *ICCV*, 2011.

[20] J. Kwon and K. M. Lee. Wang-landau monte carlo-based tracking methods for abrupt motions. *PAMI*, 35(4):1011–1024, 2013.

[21] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *PAMI*, 30(10):1683–1698, 2008.

[22] F. Liang, C. Liu, and R. J. Carroll. Stochastic approximation in monte carlo computation. *J. Amer. Statist.*, 102(477):305–320, 2007.

[23] L. Matthews, T. Ishikawa, and S. Baker. The template update problem. *PAMI*, 26(6):810–815, 2004.

[24] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai. Minimum error bounded efficient l1 tracker with occlusion detection. *CVPR*, 2011.

[25] S. M. S. Nejhum, J. Ho, and M.-H. Yang. Visual tracking with histograms and articulating blocks. *CVPR*, 2008.

[26] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. *CVPR*, 2012.

[27] D. W. Park, J. Kwon, and K. M. Lee. Robust visual tracking using autoregressive hidden markov model. *CVPR*, 2012.

[28] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. *ECCV*, 2002.

[29] D. A. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1):125–141, 2008.

[30] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. Prost: Parallel robust online simple tracking. *CVPR*, 2010.

[31] L. Sevilla-Lara and E. Learned-Miller. Distribution fields for tracking. *CVPR*, 2012.

[32] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. *CVPR*, 2005.

[33] S. Stalder, H. Grabner, and L. V. Gool. Cascaded confidence filtering for improved tracking-by-detection. *ECCV*, 2010.

[34] S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel tracking. *ICCV*, 2011.

[35] M. Yang and Y. Wu. Tracking non-stationary appearances and dynamic feature selection. *CVPR*, 2005.

[36] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), 2006.

[37] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. *CVPR*, 2012.