Robust Object Tracking via Inertial Potential based Mean Shift

Xin Sun, Hongxun Yao, Shengping Zhang
School of Computer Science and Technology
Harbin Institute of Technology
92 West Dazhi Street, Harbin 150001, China
{sunxin, H.Yao, S.Zhang}hit.edu.cn

ABSTRACT

We present a novel mean shift approach in this paper for robust object tracking based on an inertial potential model. Conventional mean shift based trackers exploit only appearance information of observation to determine the target location which usually cannot effectively distinguish the foreground from background in complex scenes. In contrast, by constructing the inertial potential model, the proposed algorithm makes good use of motion information of previous frames adaptively to track the target. Then the probability of all candidates is modeled by considering both the photometric and motion cues in a Bayesian manner, leading the mean shift vector finally converge to the location with maximum likelihood of being the target. Experimental results on several challenging video sequences have verified that the proposed method is compared very robust and effective with the traditional mean shift based trackers in many complicated scenes.

Categories and Subject Descriptors

I.2.10 [**Image Processing and Computer Vision**]: Scene Analysis—*tracking*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*motion*

General Terms

Algorithms, Experimentation

Keywords

Object tracking, mean shift, motion modeling, probabilistic approximation

1. INTRODUCTION

Object tracking is a challenging research topic in the field of computer vision and has been widely used in many applications such as surveillance, human-computer interfaces, vision-based control, and so on. In the previous literature, a huge number of tracking algorithms have been proposed, most of which search for the target in new frames with several key components: the first is object representation, which is supported to provide information of good foreground/background discrimination; then a similarity or a cost function is used to measure between the reference model and candidate targets; finally, a local mode-seeking method is needed for finding the most likely location in arriving frames.

The mean shift hill climbing method as one of the most common methods has been popular for years. After its introduction in the literature [6], it has been adopted to solve various computer vision problems, such as segmentation [2] and object tracking [4]. The popularity of the mean shift method is due its ease of implementation, real time response and robust tracking performance. The original mean shift tracker [4] uses color histograms as an object representation and Bhattacharya coefficient as a similarity measure. An isotropic kernel is used as a spatial mask to smooth a histogram-based appearance similarity function between model and target candidate regions. The mean shift tracker climbs to a local mode of this smooth similarity surface to compute the translational offset of the target blob in each frame.

Despite its promising performance, it is still a challenging problem of real-world object tracking due to variations of lighting condition, pose, scale, and view-point over time. As a result, representation of the target is of most importance for robust object tracking. However, it is exceptionally difficult to construct appearance model with respect to all of those variations in advance.

Since many tracking algorithms [4, 5, 8] based on a fixed target model are unable to track over long time intervals, some efforts have been made to well model the target to adapt to appearance changes. In [7], the authors use a patch-based dynamic appearance model in junction with an adaptive Basin Hopping Monte Carlo sampling method to successfully track a non-rigid object. Ross in [9] presents an adaptive tracking method which utilizes the incremental principal component analysis and shows robustness to large changes in pose, scale, and illumination. In [1], the authors develop an on-line feature ranking mechanism and embed it in a tracking system that adaptively selects the top-ranked discriminative features to improve tracking performance. However, few of them consider the motion information and seize the inertial property to do accurate localization.

In this paper, we present a novel mean shift approach for

ICIMCS '11, August 5-7, 2011, Chengdu, Sichuan, China

robust object tracking based on an inertial potential model. Instead of exploiting only appearance information of observation to determine the target location, the proposed algorithm, by constructing the inertial potential model, makes good use of motion information of previous frames adaptively to track the target under weak target/background distinction. Then the probability of all candidates is modeled by considering both the photometric and motion cues in a Bayesian manner, leading the mean shift vector finally converge to the location with maximum likelihood of being the target.

The rest of this paper is organized as follows: We briefly go over the mean shift framework in Section 2. In Section 3, the proposed inertial potential based tracking algorithm is described in detail. Experimental results on different challenging video sequences are shown in Section 4, and Section 5 is devoted to conclusion.

2. THE BASIC MEAN SHIFT

The mean shift method iteratively computes the closest mode of a sample distribution starting from a hypothesized mode. In specifically, considering a probability density function $f(\mathbf{x})$, given n sample points $\mathbf{x}_i$, $i = 1, \cdots, n$, in d-dimensional space, the kernel density estimation (also known as Parzen window estimate) of function $f(\mathbf{x})$ can be written as

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)}{h^d \sum_{i=1}^{n} w(\mathbf{x}_i)} \tag{1}$$

where $w(\mathbf{x}_i) \geq 0$ is the weight of the sample $\mathbf{x}_i$, and $K(\mathbf{x})$ is a radially symmetric kernel satisfying $\int k(x)dx = 1$. The bandwidth h defines the scale in which the samples are considered for the probability density estimation.

Then the point with the highest probability density in current scale h can be calculated by mean shift method as follow:

$$m_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^{n} G(\frac{\mathbf{x}_i - \mathbf{x}}{h}) w(\mathbf{x}_i)} \tag{2}$$

where the kernel profile $k(x)$ and $g(x)$ have the relationship of $g(x) = -k'(x)$.

The kernel is recursively moved from the current location $\mathbf{x}$ to the new location $m_h(\mathbf{x})$ according to mean shift vector, and finally converge to the nearest mode.

In the context of tracking, a sample corresponds to a pixel $\mathbf{x}$ and has an associated sample weight $w(\mathbf{x})$, which defines how likely the pixel $\mathbf{x}$ belongs to an object. Given the initial object position, the mean shift tracking method evaluates the new object position by computing the mean shift vector iteratively according to the equation (2). The bandwidth h defines the scale of the target candidate, i.e., the number of pixels considered in the localization process.

3. THE PROPOSED METHOD

3.1 Bayesian Formulation

Let $I_k : \mathbf{x} \rightarrow \mathbb{R}^m$ denote the image at time k that maps a pixel $\mathbf{x} = [x \ \ y]^T \in \mathbb{R}^2$ to a value, where the value is a scalar in the case of a grayscale image (m = 1) or a three-element vector for an RGB image (m = 3). Effective image preprocessing technical could also be used to generate the value. Given all the observations $I_{0:k}$ up to time k and the previous target location $\mathbf{X}_{0:k-1}$, we model the probability of the position $\mathbf{x}_i$ at time k by considering both the photometric and motion information in a Bayesian manner as

$$p(\mathbf{x}_i | I_{0:k}, \mathbf{X}_{0:k-1}) \propto \underbrace{p_p(I_k|\mathbf{x}_i)}_{photometric} \underbrace{p_m(\mathbf{x}_i|\mathbf{X}_{0:k-1})}_{motion} \tag{3}$$

where $p_p(I_k|\mathbf{x}_i)$ is the observation model that measure the similarity between the observation at the candidate state and the reference model, and $p_m(\mathbf{x}_i|\mathbf{X}_{0:k-1})$ is the motion model that presents the consistency of the current moving trend with previous.

3.2 The Photometric Model

The photometric likelihood model is still constructed due to its availability and simplicity when the feature distribution of the target is discriminative. Histograms representation of R, G, B pixel values is chosen as in [4] since it is relatively insensitive to variations of viewpoint, occlusions and non-rigidity. By normalizing their kernel weighted histograms, we can get a discrete probability density $q(j)$ for the target model, and density $p^{\mathbf{x}_0}(j)$ for the current location $\mathbf{x}_0$, where index j ranges from 1 to b, the number of histogram buckets.

Then the photometric based probability $p_p(I_k|\mathbf{x}_i)$ can be computed as

$$\mathtt{log}(p_p(I_k|\mathbf{x}_i)) \propto \sum_{j=1}^{b} \sqrt{\frac{q(j)}{p^{\mathbf{x}_0}(j)}} \delta[c(\mathbf{x}_i) - j] \tag{4}$$

where function c return the index of a pixel of its bin in the quantized feature space and δ is the Kronecker delta function.

3.3 The Inertial Potential Model

To fully exploit the motion likelihood for robust tracking, the proposed algorithm defines an inertial potential model to refine the mean shift iteration.

Dense experimental data show that, the fluctuation of the distances the target moves in two successive time slices can not be very large and usually concentrate in a small range which is independent on the target moving speed. Based on this truth, we build an inertial potential model, which presents the consistency of the current moving trend with that of the target in previous time and can update adaptively with time.

First, we calculate a reference distance according to the locations of the target in previous two frames

$$TreDis_k = (\mathbf{X}_{k-1}^x - \mathbf{X}_{k-2}^x)^2 + (\mathbf{X}_{k-1}^y - \mathbf{X}_{k-2}^y)^2 \tag{5}$$

where $\mathbf{X}_{k-1}^x$, $\mathbf{X}_{k-1}^y$ and $\mathbf{X}_{k-2}^x$, $\mathbf{X}_{k-2}^y$ denote the final locations of the target at time $k-1$ and $k-2$ respectively. The distance from a candidate position $\mathbf{x}_i$ to the last final position can be calculated as follow:

$$CanDis_i = (x_i - \mathbf{X}_{k-1}^x)^2 + (y_i - \mathbf{X}_{k-1}^y)^2 \tag{6}$$

where x_i and y_i is the coordinates of $\mathbf{x}_i$.

Then, by comparing it with the reference distance, the distance related inertial potential of pixel $\mathbf{x}_i$ can be defined as:

$$InePot_{dis}(\mathbf{x}_i) = \left\lceil \frac{|TreDis_k - CanDis_i|}{\eta} \right\rceil \tag{7}$$

where η is quantization step size and set at 500 experimentally.

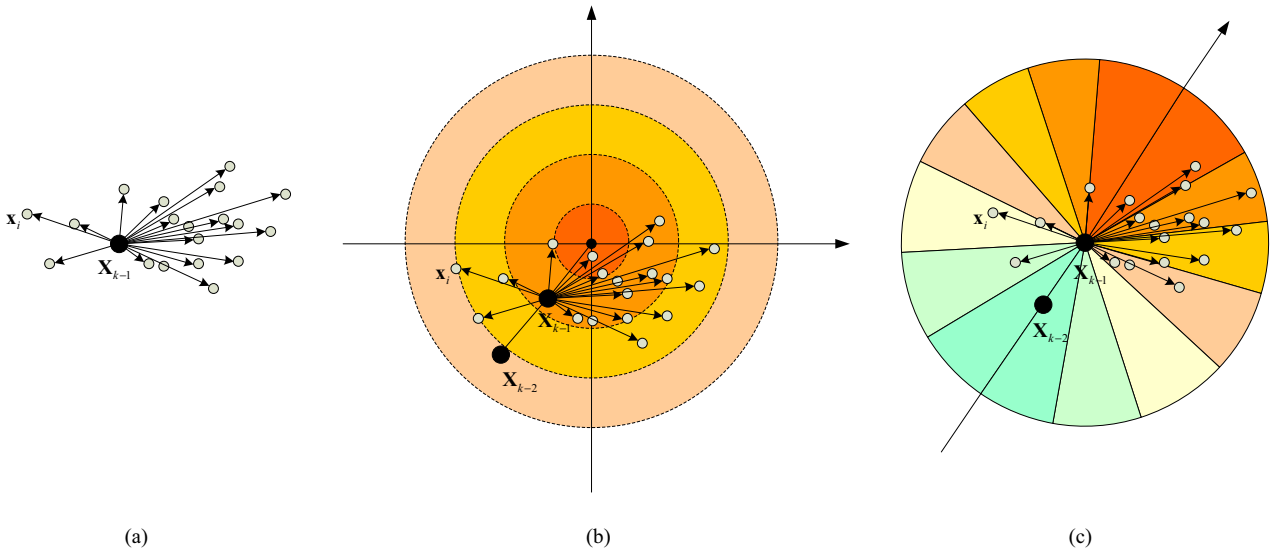

Figure 1: Illustration of the inertial potential model used in the proposed algorithm. (a) shows the initial position, derived from previous target location, and the samples in mean shift procedure. (b) and (c) present the distance and direction based inertial potential model respectively. Samples in deep orange region, maintaining the maximum consistency of moving trend with that of the target in previous time, get the largest inertial potential and vice versa for samples in light (or opposite) colored region.

Besides, the directions the object moves toward in two successive frames are usually the same or have little difference. Depending on this statistics, we can define the direction related inertial potential for pixel $\mathbf{x}_i$ by comparing the angle between the displacement of $\mathbf{x}_i$ relative to the target position in last frame with that the target moved in previous time slice:

$$InePot_{dir}(\mathbf{x}_i) = -\frac{TreDis_k + CanDis_i - CanTwoDis_i}{2\sqrt{TreDis_k \times CanDis_i}} \tag{8}$$

where $CanTwoDis_i$ is the square of the distance from $\mathbf{x}_i$ to the final target position of frame $k-2$:

$$CanTwoDis_i = (x_i - \mathbf{X}_{k-2}^x)^2 + (y_i - \mathbf{X}_{k-2}^y)^2 \tag{9}$$

Based on the description above, the motion based probability $p_m(\mathbf{x}_i|\mathbf{X}_{0:k-1})$ can be decomposed as

$$p_m(\mathbf{x}_i|\mathbf{X}_{0:k-1}) \propto p_{dis}(\mathbf{x}_i|\mathbf{X}_{0:k-1}) p_{dir}(\mathbf{x}_i|\mathbf{X}_{0:k-1}) \tag{10}$$

where

$$\log(p_*(\mathbf{x}_i|\mathbf{X}_{0:k-1})) \propto InePot_*(\mathbf{x}_i) \tag{11}$$

where $* \in \{dis, dir\}$.

Fig.1 illustrates the mechanism of the proposed inertial potential model.

3.4 Compute the Mean Shift Vector

Given the samples being weighted by $w(\mathbf{x}_i)$, we can evaluate the translation of the object centroid by computing the mean shift vector $\Delta\mathbf{x}$, such that $\mathbf{x}_1 = \mathbf{x}_0 + \Delta\mathbf{x}$. The mean shift vector is computed using the following:

$$\Delta\mathbf{x} = \frac{\sum_{i=1}^{n_h} g(\|\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\|^2) w(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_0)}{\sum_{i=1}^{n_h} g(\|\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\|^2) w(\mathbf{x}_i)} \tag{12}$$

where the weight at pixel $\mathbf{x}_i$ is specified by:

$$w(\mathbf{x}_i) = \frac{\mathtt{log}(p(\mathbf{x}_i|I_{0:k}, \mathbf{X}_{0:k-1}))}{\sum_{i=1}^{n_h} \mathtt{log}(p(\mathbf{x}_i|I_{0:k}, \mathbf{X}_{0:k-1}))} \tag{13}$$

4. EXPERIMENTAL RESULTS

In this section, we use several challenging video sequences to illustrate the advantage of our proposed algorithm. In all cases, we use the kernel with Epanechnikov profile [3] and the initial target region of the first frame was located by a rectangle supplied manually while the subsequent ones were fed by the results of previous frame. Motion compensation is used to unified the coordinate in camera moving cases.

First, we compare the proposed algorithm with the standard mean shift based on only appearance model, R, G, B histogram, on two video sequences which correspond to different challenges for visual tracking. To give a convincing comparison, experiments of the two algorithms are carried out under the same conditions. The first sequence consists of 310 frames and describes a ship, with similar color distribution to the water, navigating on the river with moving waves behind and illumination changes. Tracking results of these two algorithms are shown in Fig.2. As we can see, it is a challenge for standard mean shift tracker since it only exploit appearance information which does not provide good forground/background discrimination in complex scene where the target has the same feature property with the background region. Drafting occurs when the background pollution passes down to the following frames. In contrast, the proposed algorithm, effectively making use of motion information by constructing inertial potential model which refines the final decision, performs well.

The second video sequence describes a high jump match, which contains a player with fast and drastic motion. Cluttered background behind results in a lot of interference. Fig. 3 gives the tracking results of these two algorithms. As we

(a) The standard mean shift

(b) The proposed method

Figure 2: Tracking results of the proposed algorithm on *boat* sequence for frames of 0, 77, 110, 180, 287.

(a) The standard mean shift

(b) The proposed method

Figure 3: Tracking results of the proposed algorithm on *jumping* sequence for frames of 0, 14, 32, 53, 66.

can see, it is difficult for traditional method to capture the target while the proposed method can achieve accurate performance by exploiting motion information.

Next, we use another three sequences with different challenges to further evaluate the proposed method. The first sequence records a diving process, where dramatic appearance changes as she giving the show and undergoing shape deformation. The second video sequence shows a distance view of a parking, which contains a car driving through the square and then turn around. The background is very cluster and has lots of other stuff. The car is occluded as it passes through the lamp post. The third test is on a gray scale sequence, where a toy dog is held and swayed under a lamp. As the toy dog moves and turns, appearance and illumination changes occur. Fig.4 shows the tracking results of these three sequences, indicating the robustness of the proposed method in dealing with these challenging cases and its availability under weak target/background distinction.

5. CONCLUSIONS

A novel mean shift approach for robust object tracking based on an inertial potential model has been presented in this paper. In contrast with conventional mean shift based trackers which exploit only appearance information to determine the target location, the proposed algorithm makes good use of motion information adaptively by constructing a inertial potential model. The probability of all candidates is modeled by considering both the photometric and motion cues in a Bayesian manner, leading the mean shift vector finally converge to the location with maximum likelihood of being the target, as well as do accurate tracking under weak target/backgorund distinction. Experimental results have confirmed the effectiveness and robustness of our method.

(a) Tracking results of the proposed algorithm on *diving* sequence for frames of 50, 98, 143, 179, 190.

(b) Tracking results of the proposed algorithm on *car* sequence for frames of 2987, 3041, 3092, 3159, 3220.

(c) Tracking results of the proposed algorithm on *gray dog* sequence for frames of 295, 361, 441, 532, 611.

Figure 4: Experimental results for further evaluation.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 61071180) and NEC cooperative project (No. LC04-20101201-03).

7. REFERENCES

[1] R. Collins, L. Yanxi, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1631–1643, 2005.

[2] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *IEEE Conference on ICCV*, pages 1197–1203, 1999.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

[5] A. Elgammal, R. Duraiswami, and L. Davis. Probability tracking in joint feature-spatial spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I–781–I–788, June 2003.

[6] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[7] J. Kwon and K. M. Lee. Tracking of a non-rigid object via patchbased dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *IEEE Conference on CVPR*, pages 1208–1215, 2009.

[8] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *IEEE Conference on European Conference on Computer Vision*, pages 661–675. Copenhagen, Denmark, 2002.

[9] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *IEEE Transaction on International Journal of Computer Vision*, 77(1-3):125–141, 2008.