SALIENCY SELECTION FOR ROBUST VISUAL TRACKING

Qing Wang, Feng Chen, Wenli Xu

Tsinghua National Laboratory for Information Science and Technology
Department of Automation, Tsinghua University, Beijing 100084, P.R.China

ABSTRACT

This paper proposes a robust visual tracking approach based on saliency selection. In this method, salient patches and their spatial context inside the object region are exploited for object representation and appearance modeling. Tracking is then implemented by a hybrid stochastic and deterministic mechanism, which needs a small number of samples for particle filtering and escapes local minimum in conventional deterministic tracking. As time progresses, the selected salient patches and their spatial context are updated online to adapt the appearance model to both object and environmental changes. We carry out experiments on several challenging sequences and compare our method with the state-of-the-art algorithm to show its improvement in terms of tracking performance.

Index Terms— Saliency selection, hybrid of stochastic and deterministic tracking, adaptive appearance modeling.

1. INTRODUCTION

Visual tracking is important in many applications such as intelligent surveillance and human-computer interfaces. Persistent tracking in unconstrained environments is still challenging due to the object and environmental changes, such as pose, scale changes, deformation, camera motion and occlusion.

Many tracking methods have been proposed, most of which can be roughly categorized into two classes: deterministic tracking [1] and stochastic tracking [2]. Deterministic tracking is usually efficient, but often falls into local minimum. Stochastic tracking improves robustness over the deterministic method by its capability for escaping local minimum. However, stochastic tracking in general has a higher computational cost. In these two main tracking frameworks, object representation is the key to good tracking performance and has been emphasized in order to overcome the object and environmental variation problem during tracking.

There are two general approaches for object representation. One is to find invariant features to the changes [3], and the other is to learn appearance models online [4, 5, 6]. Since the visual invariants are difficult to obtain, the second category is more flexible, because the utilized appearance models are adaptive. The appearance model can be designed to model the object only [4], or model both the object and the background [5, 6]. Most of the latter approaches train classifiers online and have the capability to select global salient features for object representation, thereby generating superior results. However, they usually need a large feature pool and correctly labeled samples for classifier training, which are often difficult for a real tracking task.

Even in a simple global feature space, the object may have different saliency levels at different locations. Considering the human vision perception is selective [7], more salient patches should attract more attention. Inspired by this idea, we propose a tracking approach based on saliency selection. In this approach, we construct an object appearance model online by some selected salient patches and their spatial context. The object state is estimated by a hybrid stochastic and deterministic method, which needs fewer number of samples for particle filtering, and avoids the local minimum problem in deterministic tracking. Then with the tracking result, the object appearance model is updated to adapt to both the object and environmental changes.

2. SALIENCY SELECTION FOR TRACKING

We propose a saliency selection method to find salient patches for object tracking. The object state at time \( t \) is represented by \( s_t = \{ x_t, y_t, w_t, h_t, x_i^{w}, y_i^{w}\}_{i=1,2,..,M} \), where \( x_t, y_t \) denote the center position of the object, \( w_t, h_t \) denote the width and height of the object. \( x_i^{w}, y_i^{w}\) is the relative position of the \( i \)th salient patch against the object center. The state model of the target is demonstrated in Fig. 1. This model preserves the global and local information of the object.

Given the observation of the object state up to time \( t \), \( z_{1:t} = \{ z_1, \ldots, z_t \} \), the object state \( s_t \) at time \( t \) can be inferred using Bayes’ theorem:

\[
p(s_t|z_{1:t}) \propto p(z_t|s_t) \int p(s_t|s_{t-1})p(s_{t-1}|z_{1:t-1})ds_{t-1}
\]

The tracking process is governed by the dynamic model \( p(s_t|s_{t-1}) \), and the observation model \( p(z_t|s_t) \) which denotes the likelihood of \( s_t \) generating observation \( z_t \).
2.1. Hybrid Tracking Method

Note that the solution space gets large quickly as the dimension of the target state increases, so the conventional particle filter method [2] is not efficient to approximate the integration defined by Eq. 1. To overcome this difficulty, we present a hybrid method, which makes use of both the merits of stochastic and deterministic tracking approaches. Like the conventional particle filter method, this proposed method also consists of a prediction step and a correction step. In a new tracking frame, we first use the dynamic model to predict object state. And then we utilize a local mode-seeking method to obtain the final tracking result. The most advantage of this hybrid method is that it needs fewer number of particles for prediction than conventional particle filter methods, and escapes local minimum in deterministic methods.

In the prediction step, we approximate the object state \( s_t \) with \( N_s \) samples. The motion of global parameters \( \{x_t, y_t, w_t, h_t\} \) between two consecutive frames are modeled independently with Gaussian distributions, respectively. If we denote \( s^k_t = \{x_t, y_t, w_t, h_t\} \), then \( p(s^k_t|s_{t-1}^k) \) can be formulated as

\[
p(s^k_t|s_{t-1}^k) = \mathcal{N}(s^k_t; s_{t-1}^k, \Sigma)
\]

where \( \Sigma \) is a diagonal covariance matrix whose elements are the corresponding variances of the parameters, i.e., \( \sigma_{x_t}^2, \sigma_{y_t}^2, \sigma_{w_t}^2, \sigma_{h_t}^2 \). Note that the spatial context of the local salient patches is fixed in the prediction process. Therefore, the prediction of \( s_t = \{x_t, y_t, w_t, h_t, \mathbf{x}_i|_{i=1,2,...,M}\} \) can be easily obtained after the prediction of \( s^k_t \).

In the correction step, we utilize a local mode-seeking method to find the final tracking result with the predicted samples. Most performance in this hybrid method comes from the new correction step. In this step, the observation model is formulated by the likelihood at the local minimum of each patch. For each predicted sample, the initial estimate of each local patch is obtained at the prediction step, and the corresponding local minimum can be achieved by some deterministic tracking methods. In our method, we choose kernel tracker [1] for local mode-seeking. If the final mode-seeking result of some predicted sample goes out of the sample’s bounding box, we will modify the sample so that it can cover all its mode-seeking results. The modification includes moving and resizing the bounding box, which adjusts the global parameters \( \{x_t, y_t, w_t, h_t\} \) of the corresponding sample. After that, the overall observation model is defined by

\[
p(z_i|s_{t(i)}) = \prod_{i=1}^{M} \omega_i p(z_i|s_{t(i)})
\]

where \( n = 1, 2, \ldots, N_s \) means the \( n \)th sample, which may be an originally predicted sample or modified sample. \( \omega_i \) is the weight of the \( i \)th patch in the reference model, and \( p(z_i|s_{t(i)}) \) measures the similarity between the \( i \)th final local minimum in the \( n \)th sample and the reference model. The local mode-seeking is illustrated in Fig. 2. After the prediction and correction steps, the final tracking result \( \hat{s}_t \) is estimated by the Maximum a Posterior (MAP) method.

\[
\hat{s}_t = \arg \max_{s_{t(i)}} p(s_{t(i)}|z_{1:t})
\]

2.2. Online Appearance Modeling

We propose an online object appearance modeling method by saliency selection. Although many interest point/region detectors have been proposed [8], most of them treat every position in the input image equally and have high computational costs. Here we present an efficient saliency selection method, in which saliency is directly related to the tracking task. In our method, only the patches inside the object region are considered, and saliency is measured by the distance between the probability distributions of a patch and its corresponding background. A larger distance means a higher saliency level.

2.2.1. Initialization of the Appearance Model

We utilize grey-level histogram to approximate the probability distributions of local patches and the background, and use Bhattacharyya distance as the criterion for saliency selection. The Bhattacharyya distance between a local patch inside the object region and the background is defined as

\[
d(x) = \sqrt{1 - \rho[p(x), q]}
\]
where $\mathbf{x}$ is the location of the local patch. $\rho(p(x), q)$ are the histograms of a local patch and background, respectively. We choose the Bhattacharyya coefficient $\rho(p(x), q)$ as

$$\rho(p(x), q) = \sum_{u=1}^{n_b} \sqrt{p(x)_u \sqrt{q_u}}$$

where $u = 1, \ldots, n_b$ index the $u$th bin in the histogram.

To initialize the appearance model, we manually select a bounding box around the object and a background box which is 3 times as large as the object box (see Fig. 2). Considering the computational complexity, we use integral histogram [9] and search salient patches at each location inside the object region with a fixed scale. We compute the Bhattacharyya distance between each patch candidate and the background, and select the first $M$ patches with largest distances. After this, the object appearance can be modeled as $A = \{T_i, \omega_i, \mathbf{x}_i^t\}_{i=1}^{M}$, where $T_i$ and $\mathbf{x}_i$ are the kernel histogram (used for kernel local mode-seeking) and spatial context of the $i$th salient patch, respectively. The weight $\omega_i$ is proportional to the saliency level of the $i$th patch, and is normalized such that $\sum_{i=1}^{M} \omega_i = 1$.

2.2.2. Observation Model

Applying the appearance model, the likelihood defined in Eq. 3 can be computed as (we omit $t$ and $n$ without confusion)

$$p(z|x) = \prod_{i=1}^{M} \omega_i \rho^i$$

where $\rho^i$ is the Bhattacharyya coefficient between the $i$th patch in the appearance model and its corresponding local mode-seeking result.

2.2.3. Update of the Appearance Model

As time progresses, the appearance of an object may change due to many factors as discussed above. Then the initialized salient patches will be no longer salient and the appearance model will be not accurate. Therefore, adapting the appearance model to reflect these changes is crucial for long-time tracking.

After we get the tracking result in current frame, the saliency levels of local patches in the appearance model are first evaluated by their corresponding local mode-seeking results using Eq. 5. Then in the object region except the mode-seeking results, we utilize our saliency selection method to find another most salient patch. If this patch is more salient than the most non-salient patch in the appearance model, we will use the former patch to replace the latter, and then update the whole appearance model. If no such new salient patch is found, we will just update the current components in the appearance model. The kernel histogram of the salient patch is updated by

$$T_i^t = \alpha T_{i-1}^t + (1 - \alpha) T_i^{\text{new}}$$

where $T_{i-1}^t$ is the $i$th histogram in the appearance model at time $t - 1$, and $T_i^{\text{new}}$ is the histogram of the corresponding local mode-seeking result at time $t$. $\alpha \in [0, 1]$ is a forgetting factor. The spatial context is updated by

$$\mathbf{x}_i^t = \alpha \mathbf{x}_i^{t-1} + (1 - \alpha) \mathbf{x}_i^{\text{new}}$$

Finally, the weight of each patch will be normalized based on its saliency level.

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed algorithm, we test our algorithm on several challenging videos sequences. The challenging factors include occlusion, pose and scale changes, drastic deformation, unknown camera motion, etc. The number of particles is assigned as 100, and the update rate $\alpha$ is set to be 0.8.

We first utilize two experiments to demonstrate the validity of our proposed tracker, in which $M$ is set to be 10. In the first row of Fig. 3, the object undergoes long-time partial occlusion. In the second row, there are drastic pose changes of the object, and the contrast between the object and background is low. From these results, we can see that our method handles partial occlusion and pose changes well.

In the second experiment ($M$ is set to be 8), we compare the proposed algorithm with the method in [5] (referred to as VRT here). VRT selects global salient features for tracking from a large feature pool. In the NBA sequence shown in Fig. 4, the tracking player undergoes large deformation, quick movement, pose variation and severe occlusion by other players. The unknown camera motion and very similar clothes on some other players also increase the tracking difficulty. Form the tracking results, we can see our method succeeds in tracking the target all through this sequence while VRT drifts gradually from the target and fails after an occlusion. It is because the global features used by VRT are all not discriminative enough in this scene. In contrast, our approach selects local salient patches for appearance modeling, which makes our tracker more discriminative. For quantitative comparison, we plot the distances between the tracking results of these two trackers and the real target locations (labeled manually), which is shown in Fig. 5. A smaller distance implies a better method. From Fig. 4 and Fig. 5, we can conclude that our approach outperforms the conventional discriminant tracking method in handling such challenging scenarios.

4. CONCLUSION

In this paper, we propose an adaptive tracking approach based on saliency selection. In this tracking method, an appearance model is constructed online by salient selection, which preserves the spatial context of the object and can be adapted to both the object and environmental changes. Tracking is
implemented by a hybrid stochastic and deterministic mechanism, which only needs a small number of samples for particle filtering, and escapes local minimum in conventional deterministic tracking. Experiments have demonstrated the effectiveness of the proposed tracking method.

5. REFERENCES


