ROBUST BAYESIAN TRACKING ON RIEMANNIAN MANIFOLDS VIA FRAGMENTS-BASED REPRESENTATION

Yi Wu, Jinqiao Wang, Hanqing Lu

Institute of Automation, Chinese Academy of Sciences, Beijing
{ywu, jqwang, luhq}@nlpr.ia.ac.cn

ABSTRACT

Recently, the covariance region descriptor [1] has been proved robust and versatile for a modest computational cost. It enables efficient fusion of different types of features. Based on the covariance descriptor and the metric on Riemannian manifolds, we develop a robust Bayesian tracking framework via fragments-based representation in this paper. In this framework, the template object is represented by multiple image fragments or patches. Every patch votes on the possible state of the object in the current frame, by comparing its covariance descriptor with the corresponding image patch model. Tracking is then led by the Bayesian state inference framework in which a particle filter is used for propagating sample distributions over time. The weight of each particle is formulated by combining the votes of the patches using a robust statistic. Further, we extend the fast covariance computation to the Bayesian tracking problem, which makes the tracking procedure more efficient. We present extensive experimental results on challenging sequences, which demonstrate the robust tracking achieved by our algorithm.

Index Terms— Particle filter, Riemannian manifolds, covariance descriptor, integral image, Bayesian tracking

1. INTRODUCTION[1]

Object tracking is a critical task in many computer vision applications such as surveillance, augmented reality and human-computer interfaces. Target representation is one of major components for a typical visual tracker. Extensive researches have been done on this topic.

Histograms have been proved to be a powerful representation for an image region. Discarding the spatial information, the color histogram is robust to the change of object pose and shape. Several successful tracking systems have been developed using color histograms [3, 4]. Recently, Stanley *et al.* [5] proposed a novel histogram named spatiogram in which each bin is spatially weighted by the mean and covariance of the locations of the pixels that contribute to that bin. Spatiogram captures not only the values of the pixels but their spatial relationships as well. To calculate the histogram efficiently, Fatih [6] proposed a fast way to extract histograms called integral histogram. When the integral histogram has been constructed, the histogram of any rectangular region can be computed efficiently independent of the region size. Recently, Mikhail *et al.* [10] proposed a novel multi-scale histogram-based search algorithm, termed the distributive histogram, which can be evaluated exhaustively in a faster and memory more efficient manner than integral histogram.

The covariance region descriptor recently proposed in [1] has been proved robust and versatile for a modest computational cost [2, 7]. The covariance matrix enables efficient fusion of different types of features with low dimensionality. An object window is represented as the covariance matrix of features; the spatial and statistical properties as well as their correlation are characterized within the same representation. The similarity between two covariance matrices is measured on Riemannian manifolds. Fatih [2] generalized the covariance descriptor to tracking problem by simply exhaustive searching in the whole image for the region that best matches the model descriptor. This maximal likelihood estimation is very time-consuming and easily runs into problems by the background clutter. Furthermore, the spatial information encoded in the covariance descriptor is soft, so it cannot handle well partial occlusions.

Improvement for such situations is one of the benefits of our proposed robust Bayesian tracking approach. Relying on the same metric to comparing two covariance descriptors, we embed it within a sequential Monte Carlo framework and extend the fragments-based representation [9] to the particle filter implementation. The sample-based filtering technique enables to track multiple posterior modes, which is the key to escape from background distraction. Through encoding hard spatial constraint, the fragments-based representation is robust to partial occlusion. Furthermore, we extend the fast covariance computation to tracking problem with the help of integral image, which makes the tracking procedure more efficient.

2. COVARIANCE DESCRIPTOR

[1] This work is partially supported by the National Natural Science Foundation of China (Grant No. 60605004 and 60833006) and Natural Science Foundation of Beijing (Grant No. 4072025).

The covariance region descriptor proposed in [1] enables efficient fusion of different types of features and its dimensionality is small. In this descriptor an object window is represented as the covariance matrix of features. The spatial and statistical properties as well as their correlation are characterized within the same representation.

Let I be the observed image, and F be the $W \times H \times d$ dimensional feature image extracted from I

$$F(x,y) = \Phi(I,x,y) \tag{1}$$

where Φ can be any mapping such as color, gradients, filter responses, etc. Let $\{z_i\}_{i=1}^N$ be the d-dimensional feature points inside a given rectangular region R of F. The region R is represented by the $d \times d$ covariance matrix of the feature points

$$Cov_R = \frac{1}{N-1}\sum_{n=1}^{N}(z_i - \mu)(z_i - \mu)^T \tag{2}$$

where N is the number of pixels in the region R. μ is the mean of the feature points.

The element (i,j) of Cov_R represents the correlation between feature i and feature j. When the extracted d-dimensional feature includes the pixel's coordinate, the covariance descriptor encodes the spatial information of features.

For the tracking issue in this paper, $F(x,y)$ is formulated as:

$$\left(x, y, R(x,y), G(x,y), B(x,y), I_x(x,y), I_y(x,y)\right) \tag{3}$$

where (x, y) is the pixel location, R, G, B are the RGB color values and I_x, I_y are the intensity derivatives. Consequently, the covariance descriptor of a color image region is a 7×7 symmetric matrix.

3. METRIC ON RIEMANNIAN MANIFOLDS

Supposing no features in the feature vector would be exactly identical, the covariance matrix is positive definite. Thus the nonsingular covariance matrix can be formulated as a connected Riemannian manifold. A manifold is locally similar to a Euclidean space. For differentiable manifolds, the derivative at a point X lies in a vector space $\mathrm{T_X}$, the tangent space at that point. Each tangent space has an inner product $\langle , \rangle_{\mathrm{X}}$ and the norm for a tangent vector is defined by $\|y\|_{\mathrm{X}}^2 = \langle y, y\rangle_{\mathrm{X}}$.

An invariant Riemannian metric on the tangent space is defined as

$$\langle y, z\rangle_{\mathrm{X}} = tr\left(\mathrm{X}^{-\frac{1}{2}} y \mathrm{X}^{-1} z \mathrm{X}^{-\frac{1}{2}}\right) \tag{4}$$

The exponential map associated to the Riemannian metric is given by

$$exp_{\mathrm{X}}(y) = \mathrm{X}^{\frac{1}{2}} exp\left(\mathrm{X}^{-\frac{1}{2}} y \mathrm{X}^{-\frac{1}{2}}\right) \mathrm{X}^{\frac{1}{2}} \tag{5}$$

The logarithm uniquely defined at all the points on the manifold is

$$log_{\mathrm{X}}(\mathrm{Y}) = \mathrm{X}^{\frac{1}{2}} log\left(\mathrm{X}^{-\frac{1}{2}} \mathrm{Y} \mathrm{X}^{-\frac{1}{2}}\right) \mathrm{X}^{\frac{1}{2}} \tag{6}$$

For a symmetric matrix, the exponential is given by

$$exp(\Sigma) = \sum_{k=0}^{\infty} \frac{\Sigma^k}{k!} = U exp(D) U^T \tag{7}$$

Similarly, the logarithm series is

$$log(\Sigma) = \sum_{k=0}^{\infty} \frac{(-1)^{k-1}(\Sigma - I)^k}{k!} = U log(D) U^T \tag{8}$$

where $\Sigma = UDU^T$ is the eigenvalue decomposition of the symmetric matrix Σ. $exp(D)$ and $log(D)$ are the diagonal matrix of the eigenvalue exponentials and logarithms respectively.

The distance between symmetric positive definite matrices is measured by [11]

$$d^2(\mathrm{X},\mathrm{Y}) = \langle log_{\mathrm{X}}(\mathrm{Y}), log_{\mathrm{X}}(\mathrm{Y})\rangle_{\mathrm{X}} = tr\left(log^2\left(\mathrm{X}^{-\frac{1}{2}} \mathrm{Y} \mathrm{X}^{-\frac{1}{2}}\right)\right) \tag{9}$$

4. BAYESIAN TRACKING

In the Bayesian perspective, object tracking can be viewed as a state estimation problem. The purpose of tracking is to estimate $p(x_t|y_{0:t})$, which stands for the distribution of target state x_t given all observations $y_{0:t}$ up to time t.

The density propagation of $p(x_t|y_{0:t})$ can be formulated by the well-known two-step recursion:

$$\text{Prediction: } p(x_t|y_{t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{t-1})\, dx_{t-1}$$

$$\text{Update: } p(x_t|y_{0:t}) \propto p(y_t|x_t)\, p(x_t|y_{t-1}) \tag{10}$$

For visual tracking problems, the recursion can be used within a sequential Monte Carlo framework where the posterior $p(x_t|y_{0:t})$ is approximated by a weighted sample set $\{x_t^n, w_t^n\}_{n=1}^{N_s}$, where $\sum_{n=1}^{Ns} w_t^n = 1$. All the particles are sampled from a proposal density $q(x_t^n|x_{t-1}^n, y_t)$. The weight associated with each particle is formulated by:

$$w_t^n \propto \frac{p(y_t|x_t^n)\, p(x_t^n|x_{t-1}^n)}{q(x_t^n|x_{t-1}^n, y_t)} w_{t-1}^n \tag{11}$$

To prevent the weights from degenerating, we resample the particles to obtain the unweighted particle set $\left\{x_t^n, \frac{1}{N_s}\right\}_{n=1}^{N_s}$.

The common choice of the proposal density is by taking $q(x_t|x_{t-1}, y_t) = p(x_t|x_{t-1})$. As a result, the weights become the local likelihood associated with each state $w_t^n \propto p(y_t|x_t^n)$. The Monte Carlo approximation of the expectation $\hat{x}_t = \frac{1}{N_s}\sum_{n=1}^{N_s} x_t^n \approx E(x_t|y_{0:t})$ is used as the state estimation at time t.

4.1. Target dynamics modeling

Our aim is to track a region of interest in the image plane. The shape of this region is defined by a rectangle. The state is defined as $\mathrm{x} = (d, s, v)$, where $d = (x, y)$ is the location, $s = (w, h)$ represents the object size and $v = (v_x, v_y)$ is the velocity.

Commonly a first-order (B=0) or second-order auto-regressive dynamics is chosen to model the state transition:

$$x_t = Ax_{t-1} + Bx_{t-2} + C\mathcal{N}(0, \Sigma) \quad (12)$$

Matrices A, B, C and Σ defining this dynamics could be learned or be set manually. Because these parameters are difficult to learn, the AR dynamics is not suitable for our case. We propose the following scheme to predict the state.

$$\begin{aligned} d_t^n &= d_{t-1}, v_t^n = 0 && u < u_0 \\ d_t^n &= d_{t-1}^n + v_{t-1}^n && u \geq u_0 \end{aligned} \quad (13)$$

where u is a random number from distribution $U(0,1)$, $u_0 \in [0,1]$, d_t^n is the location of the n^{th} particle in time t, d_{t-1} is the tracked target location in time t-1, v_t^n, v_{t-1}^n are the velocities of the n^{th} particle in time t and t-1, respectively.

After that, we add small random disturbance to the predicted state x_t^n. The benefit of this prediction technique is that it avoids the difficult estimation of the matrices in the AR dynamics and puts no constraint on the motion of targets.

4.2. Fragments-based representation

To handle the problem of partial occlusion, the template object is represented by multiple image parts or patches to reflect spatial relationships. Since the original formulation of the fragments-based representation [9] is not suitable for the particle filter implementation, we extend this representation to Bayesian tracking in this paper.

The target is represented by a template image T. Let $P_T = (dx, dy, h, w)$ be a rectangular patch in the template, where (dx, dy) is the displacement from the template center, and w and h are the width and height respectively. Let (x, y) be a hypothesis on the target's position in the current frame. Then the patch P_T defines a corresponding rectangular patch in the image $P_{I;(x,y)}$ with the center at $(x+dx, y+dy)$ and width w, height h. Given the patch P_T and the corresponding one $P_{I;(x,y)}$, the similarity between the patches is an indication of the validity of the hypothesis that the target is indeed located at (x, y). If $d(Q, P)$ is some measure of similarity between patch Q and patch P, then the vote of the patch $P_{I;(x,y)}$ for the hypothesis is

$$V_{P_T}(x, y) = d(P_{I;(x,y)}, P_T) \quad (14)$$

For the Bayesian tracking, we denote the number of patches by N_P, the patch set of the target template by $\{P_i^T\}_{i=1}^{N_P}$. Each hypothesis or particle x_t^n is partitioned to N_P patches $\{P_i^n\}_{i=1}^{N_P}$. The distance between two corresponding patches is measured by

$$v_i^n = d(P_i^T, P_i^n) \quad (15)$$

where $d(\cdot,\cdot)$ is the metric defined in (9). Thus, for each particle we have votes $\{v_i^n\}_{i=1}^{N_P}$

Now, we want to combine the votes obtained from all template patches. A simple solution is to sum the votes, the drawback of which is that an occlusion affecting even a single patch may contribute a high value to the sum at the correct position, resulting in a wrong estimate. In other words, we would like to use a robust estimator which could handle outliers resulting from occluded patches or other reasons (*e.g.* partial pose change, such as a person turns his head).

For each particle, we order the obtained votes $\{v_i^n\}_{i=1}^{N_P}$ and choose the Q'th smallest score, which is denoted by C_i^n, as the vote for the particle.

Intuitively, the parameter Q is the maximal number of patches that we always expect to yield inlier measurements. If we are sure that occlusions will always leave at least a quarter of the target visible, then we will choose Q to be 25% of the number of patches, namely, assuming that at least a quarter of the patches will be visible.

Consequently, the local likelihood is formulated as:

$$p(y_t|x_t^n) \propto exp\{-\lambda \cdot C_i^n\} \quad (16)$$

4.3. Fast computation

With the help of integral images, the covariance descriptor can be calculated efficiently. When $d(d+1)/2$ integral images are constructed, the covariance descriptor of any rectangular region can be computed independent of the region size.

In the Bayesian tracking problems, the tracked object only occupies small part in the image, as shown in Fig.1. If we compute the integral images for the whole image, many computation resources would be wasted. Observed from our experiments, more than 60% of the computation time is used to construct the integral images. Therefore, we only compute the integral images in the region which is occupied by all the particles. This technique makes the tracking procedure more efficient.

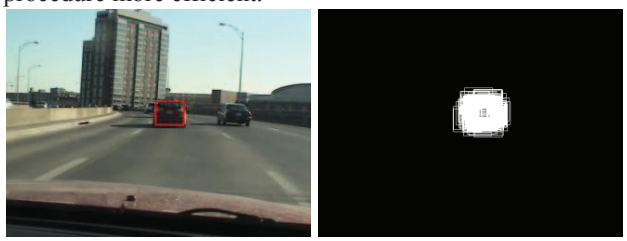


Figure 1: The state of target is shown by a red rectangle, and the corresponding particles are illustrated by white rectangles (totally 100 particles).

5. EXPERIMENTAL RESULTS

In all the experiments reported in this paper, we fixed the parameter λ of the local likelihood in (16) to the same value $\lambda = 10$. The vertical and horizontal patch structure is chosen to be the same as [9], and the template is fixed at the first frame and not updated.

The *face* sequence in Fig. 2 shows the robustness to partial occlusions. The quantitative comparison results are displayed in Fig. 3, from which we see that the tracking error of our approach is always lower than that of [2], which benefits from the fragments-based representation. The *pedestrian* sequence in Fig. 4 shows the robustness to background distraction and illumination changing. The histogram used in [9] cannot distinguish the target from the background, while the covariance descriptor can distinguish them effectively.

To illustrate the robustness against noise, we contaminated the color values with additive zero mean Gaussian noise with standard deviation $\sigma = 0.5$, where sample results are shown in Fig.4. We can see that the performance of the color-based tracker [4] significantly degrades, while our proposed approach tracks the target successfully. This owe to the average filter during covariance computation which has also been pointed out in [2].

6. CONCLUSION

Embedding the covariance-based tracker within a probabilistic framework and employing the fragments-based representation, we further improve the tracking robustness and speed. Experiments and comparisons show the robust tracking performance under partial occlusions, as well as background distraction.

The proposed Bayesian tracker is much more suitable for multi-target tracking which is our ongoing work. Due to the integral images used for fast calculation of covariance matrix, when tracking multi-object, the computational cost grows less than the linear of the tracked target number. When Covariance-based object detector [7] is used to initialize the targets, the computational cost would lower than the independent detector and tracker. This is because the detector shares the same base features (integral images) with the tracker. Furthermore, the boosted particle filter [8] can also be used to improve the multi-object tracking performance.

7. REFERENCES

[1] O. Tuzel, F. Porikli, and P. Meer. *Region covariance: A fast descriptor for detection and classification*. In ECCV 2006.

[2] F. Porikli, O. Tuzel, and P. Meer. *Covariance tracking using model update based on Lie algebra*. In CVPR 2006.

[3] D. Comaniciu, V. Ramesh, and P. Meer. *Kernel-based object tracking*. In PAMI 2003.

[4] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. *Color-Based Probabilistic Tracking*. In ECCV 2002.

[5] S. T. Birchfield, and S. Rangarajan. *Spatiograms Versus Histograms for Region-Based Tracking*. In CVPR 2005.

[6] F. Porikli. *Integral histogram: A fast way to extract histograms in Cartesian spaces*. In CVPR 2005.

[7] O. Tuzel, F. Porikli, and P. Meer. *Human Detection via Classification on Riemannian Manifolds*. In CVPR 2007.

[8] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. *A boosted particle filter: Multitarget detection and tracking*. In ECCV 2004.

[9] A. Adam, E. Rivlin and I. Shimshoni. *Robust Fragments-based Tracking using the Integral Histogram*. In CVPR 2006.

[10] M. Sizintsev, K.G. Derpanis and A. Hogue. *Histogram-Based Search: A Comparative Study*. In CVPR 2008.

[11] W. Förstner and B. Moonen. *A metric for covariance matrices*. Technical report, Dept. of Geodesy and Geoinformatics, Stuttgart University, 1999.

Frame 317 Frame 539 Frame 853

Figure 2: *face* sequence: The tracking results of [2] (row 1) and our approach (row 2).

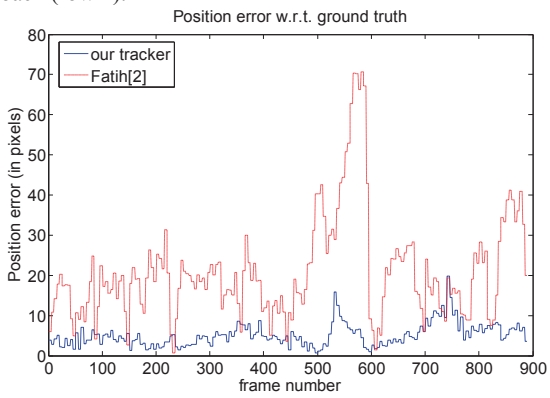
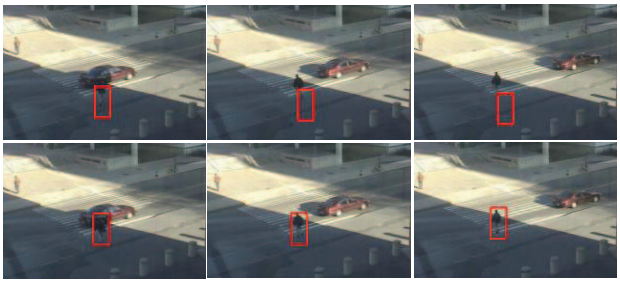
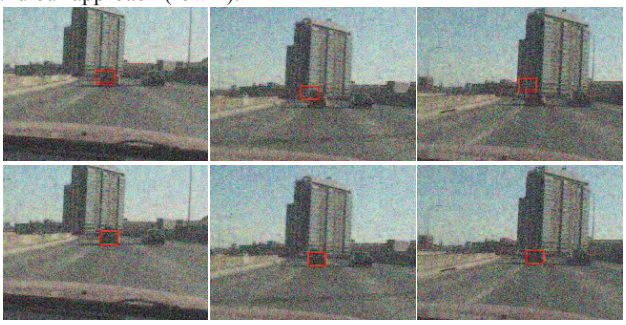


Figure 3: The quantitative comparison results of our approach and [2] on *face* sequence.



Frame 39 Frame 52 Frame 62

Figure 4: *pedestrian* sequence: The tracking results of [9] (row 1) and our approach (row 2).



Frame 29 Frame 61 Frame 89

Figure 5: *car* sequence: The tracking results of [4] (row 1) and our approach (row 2).