AFFINE MODELING OF TARGETS IN VIDEO SEQUENCES BY

PARTICLE FILTERS

Kazuhiko Kawamoto, Kyushu Institute of Technology, Japan
kawa@mns.kyutech.ac.jp

ABSTRACT

We propose an affine template matching with a statistical approach, namely particle filtering, for tracking objects of interest in video sequences. The widely used Kalman filter can not directly address the dynamics with affine transformation because of nonlinearity. In contrast, particle filter are capable of dealing with nonlinear and non-Gaussian state space models using Monte Carlo approximation. Decomposing affine transformation into six geometric parameters, we naturally model visual motion of targets by a state space model. Experimental results with real video sequences are shown to evaluate the performance.

KEYWORDS: Visual Tracking, Particle Filter, Template Matching, Appearance Model

1. INTRODUCTION

Template matching [1] is used to detect a given object in an image and is one of the most basic approaches for visual tracking in a video sequence. The naive approach is to calculate correlation between a given template and a subregion in an image. This simple approach assumes that the underlying motion of objects is translation, thus failing to track the object if other transformations such as rotation occur. Alternatively, deformable template matching [2][3] is widely used as a more flexible way. The deformation is generally described by the planar affine transformation [4]. Since the transformation is defined in the six-dimensional space, the searching of all possible transformations in the affine space is computationally prohibitive. The typical approach of avoiding this time-consuming searching is gradient descent, which iteratively updates the transformation parameters until the best match is found [5][6]. This approach is implemented in variety of numerical ways such as a Gauss-Newton and a Levenberg-Marquardt approximation, because the searching problem is formulated as nonlinear optimization [7]. Unfortunately, such algorithms face the local minimum problem due to the nonlinearity of the problem.

We propose an affine template matching with a statistical approach, namely particle filtering, for tracking objects of interest in video sequences. The algorithm explicitly deal with the dynamics of targets in video sequences, while most of the deformable template matching algorithms are based on heuristic search to track targets. If the dynamics of targets is a priori available, the use of the dynamics is in general preferable to heuristic one. The proposed algorithm finds affine transformation parameters, which describe the appearance change of the object between two successive images. Since affine transformations contain the nonlinear functions, namely trigonometric functions and multiplication, linear and Gaussian state space algorithms, namely Kalman filters [8], can not directly address the dynamics of the transformations. Thus the proposed algorithm is implemented with a particle filter (e.g. see [9]), which is capable of dealing with nonlinear and non-Gaussian state space models using Monte Carlo approximation. To evaluate the proposed algorithm empirically, an experiment with real video sequence are made.

This paper is organized as follows. Section 2 reviews particle filters for state space models. Section 3 describes geometric motion models such as affine transformations. Section 4 proposes a affine template matching tracker by particle filters. Section 5 shows experimental evaluation.
2. PARTICLE FILTERING FOR STATE SPACE MODEL

For preliminary, this section reviews particle filters for state space models [9].

The unobserved state sequence, \( x_k \in \mathbb{R}^{n_x}, k = 0, 1, \ldots, t \), is assumed to be Markovian with initial distribution \( p(x_0) \) and transition distribution \( p(x_k|x_{k-1}) \). The observations \( y_k \in \mathbb{R}^{n_y}, k = 1, 2, \ldots, t \), are conditionally independent of distribution \( p(y_k|x_k) \) given the state sequence. The general aim of time series filtering is to estimate recursively in time the posterior distribution \( p(x_t|y_{1:t}) \). This decomposition leads to sequential importance sampling [11], which updates recursively in time the importance weight as

\[
\tilde{w}_t^{(i)} \propto \frac{p(y_t|x_t^{(i)})p(x_t|x_{t-1}^{(i)})}{p(x_t^{(i)}|x_{0:t-1}, y_{1:t})}.
\] (5)

However, the importance function of the form Eq. (4) causes a degeneracy phenomenon in which almost all of the normalized importance weight in eq. (5) are close to zero after a few

\[
p(x_{0:t}|y_{1:t}) = \frac{p(x_{0:t-1}|y_{1:t-1})p(y_t|x_t)p(x_t|x_{t-1})}{p(y_t|y_{0:t-1})}.
\] (1)

We focuses on estimating the marginal distribution of the posterior distribution \( p(x_{0:t}|y_{1:t}) \), also called the filtering distribution. \( p(x_t|y_{1:t}) = \int p(x_{0:t}|y_{1:t})dx_{0:t-1} \), because visual tracking is usually interested in the current state only. Equations (1) cannot be generally computed because they require the evaluation of complex high-dimensional integrals.

Particle filters are a set of simulation-based methods that approximately and recursively estimate the posterior distribution in Eq. (1) and its marginal. The basic idea is that the posterior distribution at time \( t \) is approximately represented in the pointwise form

\[
p(x_{0:t}|y_{1:t}) \approx \frac{1}{N} \sum_{i=1}^N \delta(x_{0:t} - x_{0:t}^{(i)}),
\] (2)

where \( \delta(\cdot) \) denotes the delta-Dirac function. In Eq. (2), \( x_{0:t}^{(i)} \in \mathbb{R}^{n_x}, i = 1, 2, \ldots, N \), are assumed to be \( N \) independent and identically distributed (i.i.d) random samples, also called particles, according to \( p(x_{0:t}|y_{1:t}) \).

Unfortunately, it is usually impossible to efficiently sample points from the posterior distribution \( p(x_{0:t}|y_{1:t}) \). An alternative solution is the use of importance sampling [12]. Instead of sampling from \( p(x_{0:t}|y_{1:t}) \) directly, a so-called importance function (or proposal distribution), denoted by \( \pi(x_{0:t}|y_{1:t}) \), is used to mimic random samples from \( p(x_{0:t}|y_{1:t}) \). Here assume that random samples from the importance function \( \pi(x_{0:t}|y_{1:t}) \) are more easily available and the support of \( \pi(x_{0:t}|y_{1:t}) \) includes that of \( p(x_{0:t}|y_{1:t}) \). If \( x_{0:t}^{(i)}, i = 1, 2, \ldots, N \), are i.i.d. random samples from \( \pi(x_{0:t}|y_{1:t}) \), the empirical estimate in Eq. (2) can be modified as

\[
p(x_{0:t}|y_{1:t}) \approx \sum_{i=1}^N w_t^{(i)} \delta(x_{0:t} - x_{0:t}^{(i)}), \quad w_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}}, \quad w_t^{(i)} = \frac{p(x_{0:t}|y_{1:t})}{\pi(x_{0:t}|y_{1:t})}.
\] (3)

where \( \tilde{w}_t^{(i)} \) is the normalized weight of the so-called importance weight \( w_t^{(i)} \). Furthermore, to deal with recursive estimation problems, the importance function is assumed to be decomposed as

\[
\pi(x_{0:t}|y_{1:t}) = \pi(x_0) \prod_{k=1}^t \pi(x_k|x_0:0:k-1, y_{1:k}).
\] (4)

This decomposition leads to sequential importance sampling [11], which updates recursively in time the importance weight as

\[
\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} \frac{p(y_t|x_t^{(i)})p(x_t|x_{t-1}^{(i)})}{p(x_t^{(i)}|x_{0:t-1}, y_{1:t})}.
\]
iterations To avoid this degeneracy phenomenon, resampling methods are introduced [10][11]. The basic idea is to eliminate particles whose normalized weight are small and to concentrate on particles whose normalized weight are high.

3. AFFINE MOTION MODEL AND ITS DECOMPOSITION

If the distance between an object of interest and the camera is large, the motion of the object can be approximated as an affine transformation
\[
\begin{pmatrix}
  x' \\
  y' \\
\end{pmatrix} = \begin{pmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22} \\
\end{pmatrix} \begin{pmatrix}
  x \\
  y \\
\end{pmatrix} + \begin{pmatrix}
  t_1 \\
  t_2 \\
\end{pmatrix},
\]

or in matrix form:
\[
u' = A u + t,
\]

where \( u \) and \( u' \) are a pixel at \((x, y)\) in the image and the pixel transferred by the affine transformation with \( \det A \neq 0 \) for \( A \in \mathbb{R}^{2 \times 2}, t \in \mathbb{R} \), respectively. The matrix \( A \) is equivalent to the composed effects of rotation, scaling and shearing, and the vector \( t \) causes translation.

The affine matrix \( A \) can always be decomposed as
\[
A = R(\theta) R(-\phi) D R(\phi),
\]

where \( R(\alpha) \) and \( D \) are a planar rotation by \( \alpha \) and a diagonal matrix, respectively:
\[
R(\alpha) = \begin{pmatrix}
  \cos \alpha & -\sin \alpha \\
  \sin \alpha & \cos \alpha \\
\end{pmatrix}, \quad D = \begin{pmatrix}
  \lambda_1 & 0 \\
  0 & \lambda_2 \\
\end{pmatrix}.
\]

The matrix \( R(\theta) \) in Eq. (8) causes rotational motion and the geometric effects of the other parameters \( \phi, \lambda_1, \lambda_2 \) are the non-isotropic scaling. From Eq. (8), the elements of the matrix \( A \) in Eq. (8) are written as
\[
A = \begin{pmatrix}
  \frac{\lambda_1 + \lambda_2}{2} \cos \theta + \frac{\lambda_1 - \lambda_2}{2} \cos(\theta - 2\phi) & \frac{\lambda_1 + \lambda_2}{2} \sin \theta + \frac{\lambda_1 - \lambda_2}{2} \sin(\theta - 2\phi) \\
  \frac{\lambda_1 + \lambda_2}{2} \sin \theta + \frac{\lambda_1 - \lambda_2}{2} \sin(\theta - 2\phi) & \frac{\lambda_1 + \lambda_2}{2} \cos \theta - \frac{\lambda_1 - \lambda_2}{2} \cos(\theta - 2\phi)
\end{pmatrix}
\]

Affine transformations include Euclidean and similarity transformations as special cases. Indeed, if \( \phi = 0 \) and \( \lambda_1 = \lambda_2 (= \lambda) \) in Eq. (10), Eq. (6) is reduced to a similarity transformation in matrix form \( u' = \lambda R(\theta) u + t \). Also, if \( \phi = 0 \) and \( \lambda = 1 \), a Euclidean transformation in matrix form \( u' = R(\theta) u + t \) is obtained.

4. AFFINE TEMPLATE MATCHING BY PARTICLE FILTERS

The goal of template matching is to find the best matched region to the template in a given image. The task can be formulated as an optimization problem:
\[
\min_{u \in T} \sum_{u \in T} (I(u') - I(u))^2,
\]

where \( I(u) \) denote an image intensity at \( u \) in the image, and \( T \) is the region that form the template, that is, the summation is done over the pixels within the region. As the region \( T \), rectangles are typically chosen. If a template deformation is modeled as an affine transformation, the formulation in eq. (11) can be rewritten by
\[
\arg\min_{A, t} \sum_{u \in T} (I(A u + t) - I(u))^2,
\]

that is, the problem is to find six affine parameters, \( A, t \), which minimize the cost function of eq. (12). Note that, because of the complex pattern of intensity variations, the problem in eq. (12)
becomes a nonlinear optimization. Thus gradient descent minimization is widely used to solve the nonlinear optimization problem [5][6][7].

We present an alternative solution for affine template matching using a particle filter. Let
\[ x_t = (\theta(t), t_1(t), t_2(t), \lambda_1(t), \lambda_2(t), \phi(t))^T \] (13)
denote a state vector consisting of six affine parameters at time \( t \). The state transition of \( x_t \) is assumed to be Markovian, i.e., \( p(x_t | x_{t-1}) \) and the initial distribution \( p(x_0) \) is assumed to be given. For the state transition, assume that the affine parameters smoothly change in a video sequence, i.e.,
\[ x_t = x_{t-1} + v, \quad v \sim N(0, \Sigma) \] (14)
where \( \Sigma = \text{diag}(\sigma_\theta^2, \sigma_{t_1}^2, \sigma_{t_2}^2, \sigma_{\lambda_1}^2, \sigma_{\lambda_2}^2) \). Thus a particle filter, as described in Sec. 2, is applied to the template matching problem. As an importance distribution in eq. (4), the prior distribution \( p(x_t | x_{t-1}^{(i)}) \) is adopted. In this case, the importance weight in eq. (5) reduces to the simple form [13][14]:
\[ \tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} p(y_t | x_{t-1}^{(i)}). \] (15)

As the observation density, the distribution
\[ p(y_t | x_{t-1}^{(i)}) \propto \exp \left( -\sum_{u \in T} \frac{|I(A_t u + t_t) - I(u)|}{\#T} \right) \] (16)
is adopted, where \( \#T \) is the number of the pixels in the template \( T \) and \( A_t, t_t \) consist of the affine parameters \( x_t \). Although the state transition in eq. (14) are linear for each parameter, the existence of the trigonometric functions in eq. (16) results in a nonlinearity filtering problem. After calculating the likelihoods, the particles \( x_{t-1}^{(i)}, i = 1, 2, \ldots, N \), are resampled with replacement according to the importance weight \( w_t^{(i)} \), \( i = 1, 2, \ldots, N \).

5. EXPERIMENTS WITH REAL SEQUENCES

Figures 1 are sample images of a traffic video sequence used for experimental evaluation. This video sequence are 768 x 576 in size and taken with a stationary camera at an intersection by the Institut für Algorithmen und Kognitive Systeme group in Universität Karlsruhe, Germany.

In this experiment, the target of interest is a car coming into range of vision from the right side. A template is used for the experiment, which is obtained by clipping out of the first image of the sequence by hand. These results are obtained setting \( \sigma_\theta = 5^\circ, \sigma_{t_1} = 7(\text{pixel}), \sigma_{t_2} = 7(\text{pixel}), \sigma_\phi = 5^\circ, \sigma_{\lambda_1} = 0.1, \sigma_{\lambda_2} = 0.1 \) and the number of particles to be 2000. Figures 1 shows the means estimated from the filtering distribution. From the results, the proposed algorithm can

Figure 1. Experimental results of a traffic scene video.
correctly track the target in the video sequence even though the appearance of the target car changes in the video sequence (turning right) and some occlusions happen.

The second experiment is made with a TV program video from the motion imagery database by the Signal Analysis and Machine Perception Laboratory, the Ohio State University, USA. Figures 2 are sample images of the video which is 352 x 288 in size and taken with a camera with pan/zoom motion. These results are obtained setting $\sigma_\theta = 3^\circ$, $\sigma_{t_1} = 3$ (pixel), $\sigma_{t_2} = 3$ (pixel), $\sigma_\phi = 3^\circ$, $\sigma_{\lambda_1} = 0.1$, $\sigma_{\lambda_2} = 0.1$ and the number of particles to be 1000. Figures 2 shows the means estimated from the filtering distribution. In the video, the camera is moving with pan/zoom motion and the face is also moving with tilt motion. In this case, a translation template matching, which is widely used, fails to track the target. Our affine modeling approach can capture these motions smoothly in the video sequence.

6. Conclusions

We formulate template matching as a time series filtering problem and proposes an affine template matching for visual tracking using particle filters. Although affine transformations are linear in pixel positions, the parameterized elements include nonlinear functions. Then Kalman filter [8], which is capable of dealing with linear and Gaussian state space models, can not be applied directly. Particle filters are a powerful tool for time series filtering because it is capable of dealing with nonlinear and non-Gaussian state space models. Consequently, particle filters are suited to affine template matching. The proposed algorithm determines the affine parameters which describe the appearance change of the object between two images. In the experiment, even though the appearance of the target car changes in the video sequence (turning right) and some occlusions happen, the proposed algorithm can correctly track the target through the sequence. The experiment with the real video sequences shows the effectiveness of the particle filter. This modeling technology gives a basis for traffic analysis such as congestion prediction accident monitoring in Intelligent Transport Systems and man–machine interaction in video analysis.

Acknowledgements

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under the Grant-in-Aid for Scientific Research, Young Scientists B (No.16700169).

References