Optik 125 (2014) 1680–1686

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Robust visual tracking base on adaptively multi-feature fusion and particle filter

Jian-fang Dou*, Jian-xun Li

Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 24 April 2013
Accepted 5 September 2013

Keywords:
Particle filter
Visual tracking
LBP
Partial occlusions
Adaptively fusion

ABSTRACT

In order to avoid the tracking failure based on single feature under the conditions of cluttered backgrounds illumination changes, a robust tracking algorithm was proposed based on adaptively multi-feature fusion and particle filter. Color histogram was used to describe the overall distribution characteristics of the target and histogram of oriented gradients containing some construction information and LBP is very effective to describe the image texture features. The Three features were fused in the frame of particle filter. Meanwhile, the weights of each feature were adjusted dynamically. The experimental results show that with adaptive fusion, the tracker becomes more robust to illumination changes, pose variations, partial occlusions, cluttered backgrounds and camera motion.

1. Introduction

Visual tracking [1] has become a popular topic in the field of computer vision. Its potential applications include smart surveillance, virtual reality, perceptual interface, video conferencing, quality inspection on a production line, video surveillance of a public building, or vision-based navigation of an autonomous robot, etc. Although visual tracking has been intensively studied in the literature, developing a robust tracking algorithm in complex environments is still an open problem.

In the literature, several algorithms can be found dealing with precise object tracking in real time. However, most of them are based on a single cue or modality and are, therefore, often limited to a particular environment that is, typically, static, controlled, and known a priori. It can be taken for granted that no single visual cue will be robust and general enough to deal successfully with the wide variety of conditions occurring in real-world scenarios. In this paper, we argue that the integration of several different cues will provide means to overcome such limitations. Combining a multitude of complementary cues will lead to an enlarged working domain of the whole system; whereas combining redundant cues will lead to an increased reliability of the system. Continuous evaluation of the individual cues will provide means for dynamic adaptation of the cue parameterization as well as of the applied integration mechanism. Optimal context-dependent fusion of information will be key to the long-term goal of robust object tracking.

In this paper, we proposed a robust visual tracking based multi-cues in particle filter framework in complex environments. Particle filter is adopted to solve the non-linear and non- Gaussian problems in visual tracking. The three cues are color reference models, gradient orientation and LBP texture. Color reference model [7,8] have in particular been proved to be very useful for tracking tasks where the objects of interest can be of any kind, and exhibit in addition drastic changes of spatial structure through the sequence, due to pose changes, partial occlusions, etc. For a better target representation, the gradient or edge features have been used in combination with color histogram [9]. The local binary pattern (LBP) [10,11] technique is very effective to describe the image texture features. LBP has advantages such as fast computation and rotation invariance, which facilitates the wide usage in the fields of texture analysis, image retrieval, face recognition, image segmentation. An adaptive multi cue integration strategy is: calculate the weight of each cue by the predicted reliability between the Euclidean distance of estimated target state through fused cue and each cue respectively, Then the weight of each frame can be determined in advance, Through this we can stably tracking object in complex conditions. The experimental results show that with adaptive fusion, the tracker becomes more robust to illumination changes, pose variations, partial occlusions, cluttered backgrounds and camera motion.

2. Related work

Visual tracking can be considered to match coherent relations of image features between frames. The last decades various tracking

* Corresponding author.
E-mail address: specialdays_2010@163.com (J.-f. Dou).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.10.007

algorithms have been proposed [2–5,13]. However, most of them are based on a single image cue. It is clear that no single image cue can be robust enough to success fully deal with various conditions occurring in the real-word scenarios. To overcome the weak robustness of single-cue tracking, many algorithms have been proposed based on multi-cue fusion. Multiple cues fusion not only can provide more reliable observation when estimating a state, but different cues may be complementary in that one may succeed when another fails. The key challenge for this kind of algorithm is how to optimally fuse multiple cues. In most algorithms [6], the fusion scheme is non-adaptive. In which the reliability of each cue is assumed to be unchanged during the tracking. Such assumption is often invalid due to the dynamically changing environments. To overcome this problem, a novel tracking method based on adaptive fusion and particle filter is proposed.

The paper is organized as follows. Section 3 briefly introduces the particle filter algorithm. Section 4 presents the robust visual tracking algorithms in detail. Experimental results are presented and discussed in Section 5. Section 6 concludes the paper.

3. Particle filter

We use the particle filter (PF) framework to guide the tracking process. PF [12] is a well known Bayesian sequential importance sampling technique used for posterior distribution estimation of state variables in a dynamic system. PF becomes a popular tracking framework since the seminal work in [13] for nonlinear, non-Gaussian environments. The framework contains mainly two iterative steps: (1) a prediction step that is used to predict the target in the current frame based on previous observations, and (2) an update step that maintains sample (particle) weights for the Bayesian inference.

In the context of visual tracking, let $\{y_1, y_2, \cdots\}$ be the observations (e.g., appearance in video frames) and $\{x_1, x_2, \cdots\}$ be the states (e.g., poses of target objects in the video), where y_t and x_t denote the observation and state variable at time t respectively. The task of prediction is to estimate the distribution of x_t given all previous observations $y_{1:t-1} = \{y_1, y_2, \cdots, y_{t-1}\}$ up to time $t-1$. This conditional distribution can be recursively computed as

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

At time t, the observation y_t is available and the state vector is updated using the Bayes rule

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \tag{2}$$

where $p(y_t|x_t)$ denotes the observation likelihood.

In the particle filter, the posterior $p(x_t|y_{1:t})$ is approximated by a finite set of N samples $\{x_t^i\}_{i=1,\cdots,N}$ with importance weights w_t^i. The candidate samples x_t^i are drawn from an importance distribution $q(x_t|x_{1:t-1}, y_{1:t})$ and the weights of the samples are updated as

$$w_t^i = w_{t-1}^i \frac{p(y_t|x_t^i)p(x_t^i|x_{t-1}^i)}{q(x_t|x_{1:t-1}, y_{1:t})} \tag{3}$$

The samples are resampled to generate a set of equally weighted particles according to their importance weights to avoid degeneracy. In the case of the bootstrap filter $q(x_t|x_{1:t-1}, y_{1:t}) = p(x_t|x_{t-1})$ and the weights become the observation likelihood $p(y_t|x_t)$.

4. Tracking algorithm with adaptively multi-feature fusion

4.1. Motion model

For the purpose of tracking objects in video sequences, we initially choose a rectangular region that defines the object. The state vector of the object region is parameterized by $X_k = [x, y, w, h]$, where (x, y) denotes the centroid of the rectangle, and w and h are the height and width, respectively. The motion model is modeled by a random walk equation

$$X_k = X_{k-1} + W_k \tag{4}$$

where W_k is the process noise and assumed to be a multivariate normal distribution with the covariance matrix $Q = diag(\sigma_x^2, \sigma_y^2, \sigma_w^2, \sigma_h^2)$, describing the uncertainty in the state vector.

4.2. Color distribution model

A target is usually defined by a rectangle or an ellipsoidal region in the image. Most existing target tracking schemes use the color histogram to represent the rectangle or ellipsoidal target. In this paper, we will present a new target representation approach by using the joint color-texture histogram. First let us review the target representation in the mean shift tracking algorithm [8].

Denote by $\{x_i^*\}_{i=1\cdots n}$ the normalized pixel positions in the target region, which is supposed to be centered at the origin point. The target model $\hat{q}$ corresponding to the target region is computed as

$$\begin{cases} \hat{q} = \{\hat{q}_u\}_{u=1\cdots m} \\ \hat{q}_u = C\sum_{i=1}^{n} k(||x_i^*||^2)\delta[b(x_i^*) - u] \end{cases} \tag{5}$$

where $\hat{q}_u$ represent the probabilities of feature u in target model $\hat{q}$, m is the number of feature spaces, δ is the Kronecker delta function, $b(x_i^*)$ associates the pixel. x_i^* to the histogram bin, $k(x)$ is an isotropic kernel profile and constant C is a normalization function defined by

$$C = 1/\sum_{i=1}^{n} k(||x_i^*||^2) \tag{6}$$

Similarly, the target candidate model $\hat{p}(y)$ corresponding to the candidate region is given by

$$\begin{cases} \hat{p}(y) = \{\hat{p}_u(y)\}_{u=1\cdots m} \\ \hat{p}_u(y) = C_h\sum_{i=1}^{n_h} k(||\frac{y - x_i}{h}||^2)\delta[b(x_i) - u] \end{cases} \tag{7}$$

$$C_h = 1/\sum_{i=1}^{n_h} k(||\frac{y - x_i}{h}||^2) \tag{8}$$

where $\hat{p}_u(y)$ represents the probability of feature u in the candidate model $\hat{}\, \hat{p}(y)$, $\{x_i\}_{i=1\cdots n_h}$ denote the pixel positions in the target candidate region centered at y, h is the bandwidth and constant C_h is a normalization function.

In order to calculate the likelihood of the target model and the candidate model, a metric based on the Bhattacharyya coefficient is defined between the two normalized histograms $\hat{p}(y)$ and $\hat{q}$ as follows:

$$\rho(\hat{p}(y), \hat{q}) = \sum_{u=1}^{m} \sqrt{\hat{p}_u(y)\hat{q}_u} \tag{9}$$