Signal Processing: *Image Communication* 29 (2014) 590–598

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Robust vision tracking by online random ferns and template library

Qingge Ji *, Peng Zhang, Jinghong Du

School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form
5 January 2014
Accepted 14 March 2014
Available online 4 April 2014

Keywords:
Vision tracking
Random ferns
Template library
Discriminative and generative models

ABSTRACT

In this paper, we proposed a robust tracking algorithm with an appearance model based on random ferns and template library. We adopt random Gaussian difference to generate binary features which depend on two randomly selected points and their corresponding Gaussian blur kernels. Semi-naive Bayes based random ferns are adopted as the discriminative model, and a template library including both positive templates and negative templates is used as generative model, the co-training of both discriminative and generative models gives our tracker the ability to separate foreground and background samples accurately. Besides, we also come up with a fragment based method which combines global ferns and local ferns to handle the occlusion problem. Experimental results demonstrated that the proposed algorithm performs well in terms of accuracy and robustness.

1. Introduction

The ability to track unknown objects in realistic environments is important for a wide range of applications, e.g. intelligent surveillance, human–computer interfaces, and robot navigation. For unknown objects, we have no prior knowledge of their appearance. Under this condition, objects may be only specified at runtime, usually at the first frame of the sequence. Although object tracking has been studied for years, tracking generic objects remains to be a challenging work. The challenges come from noise, occlusion, varying viewpoints, background clutter, illumination change, abrupt motion and so on.

Recently, a lot of tracking algorithms have been proposed to overcome these challenges. Among them, a class of tracking techniques called the tracking by detection has become particularly popular, which treats the tracking problem as detection over each frame. Tracking by detection is commonly performed using sliding window classifiers [6]. The sliding window classifier trains the discriminative model and scans over locations in the image at multiple scales, and chooses the best sub-window which has the highest score as the target. Since the object appearance changes over time, we need to model the object appearance online and gradually adapt to its current changes. There has been a lot of algorithms to model the object appearance online. Oza and Russell [13] proposed an online version of bagging and boosting, which requires only one time passing through the training data. Grabner and Bischof [9] proposed a novel online AdaBoost feature selection method and demonstrated excellent performance on the visual tracking task. Ross et al. [17] presented a tracking method that incrementally learns low-dimensional subspace representation, and efficiently adapts to the appearance changes of the target object online. Driven by high computational efficiency and excellent discriminative performance, random forest has been

* Correspondence to: Department of Computer Science, School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, P.R. China.
E-mail address: issjqg@mail.sysu.edu.cn (Q. Ji).

http://dx.doi.org/10.1016/j.image.2014.03.001
0923-5965/

used in different applications of computer vision. Saffari et al. [19] proposed a novel online random forest algorithm and demonstrated that with increasing training samples its performance converges to the off-line version, experiments on visual tracking demonstrated its real time superior performance. Zhang et al. [22] adopted a very sparse measurement matrix to efficiently extract the Haar-like features, and the tracking task was formulated as a binary classification using the naive Bayes classifier. Kalal et al. [10] presented a tracking-modeling-detection framework where the target object's appearance is robustly modeled by two processes (growing and pruning events).

However, these tracking-by-detection algorithms also have their drawbacks. These methods train a discriminative classifier online, since they bootstrap themselves by using the current state of the tracker to extract positive and negative samples, slight inaccuracy may accumulate and cause drifting. To overcome this problem, Babenko et al. [3] presented a novel online multiple instance learning algorithm for object tracking and achieved superior and real-time performance. There has been other ways to avoid the error accumulation problem. Blum and Mitchell [4] proposed a co-training framework that trains two classifiers on two conditionally independent views of the same data and then uses the evaluation from each classifier to enlarge the training set of each others. It has been proved that the co-training framework can find an accurate decision boundary starting from a small quantity of labeled data if the two feature sets are independent. Yu et al. [21] proposed a co-training based approach which continuously labels incoming data and online updates a hybrid discriminative and generative model. Results demonstrated that under challenging situations, this method has strong reacquisition ability and robustness to distracters in background.

Ozuysal et al. [16] proposed a simple but efficient and robust feature point recognition algorithm. The algorithm uses non-hierarchical structures which also refer to as ferns to classify the patches around feature points. In this paper we make some improvement of ferns for visual tracking. Instead of randomly generating the binary features at initial phase and keeping unchanged until the end of the tracking process, we propose a feedback mechanism to drop the less discriminative fern and regenerate a new one. This approach makes it possible to use only a few number of ferns and acquire high discriminative ability.

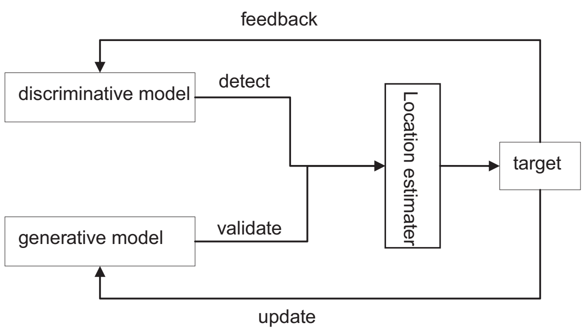

Fig. 1. Framework of our proposed algorithm.

The overall framework of our algorithm is shown in Fig. 1. In addition to using random ferns as our discriminant model, we also propose a template library as our generative model which gradually absorbs new samples of object and keeps them in the model as the system's memory. The memory is used to validate the decision made by discriminative model. The hybrid of discriminative and generative models demonstrated superior performance on challenging sequence.

2. Discriminative model: adaptive random fern

2.1. Random fern

Random ferns are similar to the random forest, which replace the trees by non-hierarchical ferns and pooling their answers in a semi-Naive Bayes. Let $c_i, i=1,\ldots,H$ is a set of classes and $f_j, j=1,\ldots,N$ is a set of binary features that extracted over the patch which to classify. We try to look for

$$\hat{c}_i = \arg\max_{c_i} P(C=c_i|f_1,f_2,\ldots,f_N). \tag{1}$$

Using Bayes's theorem,

$$P(C=c_i|f_1,f_2,\ldots,f_N) = \frac{P(f_1,f_2,\ldots,f_n|C=c_i)P(C=c_i)}{P(f_1,f_2,\ldots,f_n)}. \tag{2}$$

Since the denominator does not depend on C, we take it as constant and just ignore it. Using the notation of naive independence assumptions,

$$P(C=c_i|f_1,f_2,\ldots,f_N) = f(C=c_i)\prod_{i=1}^{n} P(f_i|C=c_i) \tag{3}$$

Completely ignoring the correlations between features always leads to a less discriminative classifier, while a complete representation of the joint probability in Eq. (1) is not feasible since it requires to store 2^N entries for each class, where N is the total number of binary features. Since binary feature is very simple, we require a lot of features ($N \geq 200$) for accurate classification. To make the problem tractable while accounting for the correlations among features, we find a compromised way to divide our feature set into M groups of size $S=N/M$. Then we compute the joint probability of features in each group and simply ignore the dependency among groups. The conditional probability in Eq. (3) becomes

$$P(C=c_i|f_1,f_2,\ldots,f_N) = P(C=c_i)\prod_{i=1}^{M} P(F_i|C=c_i) \tag{4}$$

where $F_k = \{f_{r1}, f_{r2}, \ldots, f_{rS}\}$, and $r1, r2, \ldots, rS$ are randomly selected features as a group from the feature set N. Each group of features is used to train a unique classifier which we name it as fern. A fern can be viewed as a histogram which encodes the joint probability distribution of its corresponding group of features.

It is believed that human retina extracts details from images using difference of Gaussian (DoG). We proposed a feature extraction method based on DoG. Inspired by FREAK [2], we extract our binary feature by thresholding difference between pairs of pixels with their corresponding Gaussian kernels. The binary feature f is generated by Random Gaussian Binary Test (RGBT), described as