Optik 126 (2015) 5287–5293

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Improved weighted multiple instance learning for object tracking

Jianfang Dou*, Qin Qin, Zimei Tu

Department of Automation and Mechanical and Electrical Engineering, School of Intelligent Manufacturing and Control Engineering, Shanghai Second Polytechnic University, Buildings 16-521, Shanghai 201209, China

ARTICLE INFO

Article history:
Received 25 November 2014
Accepted 26 September 2015

Keywords:
Visual tracking
Distribution field
Multiple instance learning
Tracking-by-detection

ABSTRACT

Visual tracking usually requires an object appearance model that is robust to changing illumination, partial occlusion, large pose and other factors encountered in video. Currently, a technique called "tracking by detection" has been developed and studied with promising results. A typical tracking by detection algorithm called MIL (Multiple Instance Learning) has become one of the most popular methods in tracking domain. This technique is designed to alleviate the drift problem by using an MIL based appearance model to represent training data in the form of bags. In this paper, we improved the WMIL tracker in two aspects. First, we propose a new bag model that integrates the importance of the samples in the positive bag naturally with the distribution field of the image patches as a part of the weighting function. Then, in order to solve the potential overfitting problem, we propose a dynamic function to estimate probability of instance by introducing extra parameters instead of the original logistic function. Experiments on some publicly available benchmarks of video sequences demonstrate the effectiveness and robustness of our approach.

1. Introduction

Visual tracking, one of the fundamental topics in computer vision, has long been playing a critical role in numerous applications such as surveillance, military reconnaissance, motion recognition and traffic monitoring. While much breakthrough has been made within the last decades [1,2], it still remains challenging in many aspects including pose variation, illumination change, partial occlusion, camera motion and background clutter.

Numerous tracking algorithms have been proposed in the literature over the past couple of decades and they can be categorized into two classes based on their different appearance representation schemes [3,4]: generative models and discriminative ones. Generative algorithms typically learn an appearance model to represent the object and then search for image regions with minimal reconstruction errors as tracking results for current frame. To deal with appearance variation, Ross et al. [5] proposes a tracking method that incrementally learns a low-dimensional subspace representation, efficiently adapting online to changes in the appearance of the target. The model update, based on incremental algorithms for principal component analysis, includes two important features: a method for correctly updating the sample mean, and a forgetting factor to ensure less modeling power is expended fitting older observations. Both of these features contribute measurably to improving overall tracking performance. Recently, Mei et al. [6] proposes a ℓ_1 tracker based on sparse representation where the object is modeled by a sparse linear combination of target and trivial templates and treating partial occlusion as arbitrary but sparse noise. As it essentially solves a series of ℓ_1 minimization problem, the computational complexity is too high and therefore limits its performance. Although dozens of methods have been proposed to speed up the ℓ_1 tracker, they also fail to make much improvement in the trade off accuracy and time consumption. In one word, these generative models do not take into account background information, throwing away some very useful information that can help to discriminate object from background.

On the other hand, discriminative approaches consider visual tracking as a binary classification problem, for it does not build an exact representation of the target but tries to find decision boundaries between the object and its surrounding background within a local region. Discriminative approaches utilize not only features extracted from the object but also features from the background. Collins et al. [7] have demonstrated that selecting discriminate features in an online manner can greatly improve the tracking performance. Recently, some successful tracking algorithms training their classifiers using boosting techniques have been proposed. Dietterich et al. [8] proposes a novel on-line Adaboost feature selection algorithm, depending on the background the algorithm

* Corresponding author. Tel.: +86 50217415.
E-mail address: k1882001@163.com (J. Dou).

http://dx.doi.org/10.1016/j.ijleo.2015.09.196
0030-4026/

selects the most discriminative features for tracking resulting in stable tracking results, which demonstrates to handle appearance change of the object like illumination changes or out of plane rotations naturally. However, as it utilizes the tracking result in current frame as the only one positive example to update the classifier, slightly inaccuracy of the tracking position in current frame can bring mislabeled training samples for the classifier. These errors may accumulate over time, leading to model drift or even tracking failures. Grabner et al. [9] presents a novel on-line semi-supervised boosting method, the main idea is to formulate the update process in a semi supervised fashion as combined decision of a given prior and an on-line classifier, this comes without any parameter tuning and significantly alleviates the drifting problem. The weakness of this method is that the classifier is trained by only labeling the examples at the first frame while leaving the samples at the coming frames unlabeled, which loses valuable motion information between frames.

In order to solve the similar ambiguity in face detection, Babenko et al. [10,11] suggests employing multiple instance learning approach for object tracking problem and proposes MIL tracker. Their seminal work uses an MIL based appearance model to represent training data in the form of bags. It obeys the rule that a positive bag should contain at least one positive example and examples in negative bag are all negative. The positive examples are cropped around the object while the negative ones are far from the object location. The classifier is then trained in an online manner using the bag likelihood function. Because some flexibility is allowed in separating training examples, MIL tracker can solve the inherent ambiguity by itself to some degree, leading to more robust tracking. Zhang [12] points out that the positive instance near the object location should contribute larger to the bag probability, i.e., the weights for the instance near the object is larger than that far from the object location, thus integrating the sample importance into the learning procedure and proposing a weighed MIL (WMIL) tracker which demonstrates superior performance. We observe that the Noisy-OR model used in the MIL tracker does not take into account any information about the importance of the positive samples. Therefore, MIL tracker may select less effective positive samples. Oppositely, though WMIL integrates the sample importance into the learning procedure, the bag probability made up of weighed sum of instance probability seems to weaken the discriminative power for the final strong classifier and may bring new ambiguity, i.e., the normalized weights are not sufficient enough to keep the desired property that if one of the instance in a bag has a high probability, the bag probability will be high as well. Another important issue is that the standard logistic function which transfers output of strong classifiers to probability of an instance being positive has potentially overfitting problem which degrades the discriminative model severely. For the target representation, recently, Haar-like feature [13] is widely used to represent the target in tracking-by-detection because of its fast computing by integral image and strong ability to represent the spatial structure of the target. Sevilla-Lara and Learned-Miller [14] using Distribution Fields (DFs) to build an image descriptor, able to smooth the objective function and keep the information about pixel values.

Motivated by above-mentioned discussions, in this paper, we enhance the WMIL tracker in two aspects. First, we propose a new bag model that integrates the importance of the samples in the positive bag naturally with the distribution field of the image patches as a part of the weighting function. Then, in order to solve the potential overfitting problem, we propose a dynamic function to estimate probability of instance by introducing extra parameters instead of the original logistic function.

The paper is organized as follows: In Section 2, we review the tracking method with multiple instance learning and introduce the distribution fields. The proposed method is shown in Section 3. Section 4 gives the detailed experiment setup and results. Finally, Section 5 concludes the paper.

2. Preliminaries

2.1. Multiple instance learning

In the online MIL [10,11], the training samples are expressed by the set form $\{(X_i, y_i)\}_{i=1,2,\ldots,N}$, where X_i stands for a training sample (i.e., one bag). For training instance set $\{x_{ij}\}$, the binary bag label $y_i \in \{0, 1\}$ is defined as: $y_i = \max(\{y_{ij}\})$, where $y_i \in \{0, 1\}$, are the instance labels. Therefore, the label of the bag is considered positive if it contains at least one positive instance, otherwise the bag label is negative. Online MIL uses the gradient boosting framework to train a boosting classifier that maximizes the log likelihood of bags:

$$L\left(H \middle| \left\{(X_i, y_i)\right\}_{i=1,2,\ldots,N}\right) = \sum_{i=1}^{N} \left(y_i \log\left(p\left(y_i = 1|X_i\right)\right) + \left(1 - y_i\right) \log\left(1 - p\left(y_i = 1|X_i\right)\right)\right) \quad (1)$$

Notice that the likelihood is defined over bags but not instances, because instance labels are unknown during training, and yet the goal is to train an instance classifier that estimates $p(y|x)$. We therefore need to express $p(y_i|X_i)$ the probability of a bag being positive, in terms of its instances. In [10,11] the Noisy-OR (NOR) model is adopted:

$$p\left(y_i|X_i\right) = 1 - \prod_j \left(1 - p\left(y_i|x_{ij}\right)\right) \quad (2)$$

For each instance x_{ij}, the probability of the instance being positive is calculated by:

$$p\left(y_i|x_{ij}\right) = \frac{\exp\left(H\left(x_{ij}\right)\right)}{\exp\left(H\left(x_{ij}\right)\right) + \exp\left(-H\left(x_{ij}\right)\right)} \quad (3)$$

Similar to the online boosting, the online MIL Boosting classifier is comprised by linear combination of K selectors. To update the MIL classifier $H(x)$, all the weak classifiers are updated first, and then each $Sel_t(x)$ sequentially selects a real-value weak classifier by minimizing the loss function $L(H|\{(X_i, y_i)\}_{i=1,2,\ldots,N})$, which is given by:

$$\mathrm{Sel}_t\left(x\right) = \underset{h \in H}{arg\,\min} L\left(H|\left\{(X_i, y_i)\right\}_{i=1,2,\ldots,N}\right) \quad (4)$$

where H is the strong classifier comprised by the first $t-1$ selectors. As to the real-value weak classifier h, the log odds ratio is utilized in MIL, given by:

$$h\left(x\right) = \log\left[\frac{p\left(y = 1|f_k\left(x\right)\right)}{p\left(y = 0|f_k\left(x\right)\right)}\right] \quad (5)$$

$p_k(f_k(x)|y=1) \sim N(\mu_1, \sigma_1)$ and similarly for $y=0$. We let $p(y=0) = p(y=1)$ and use Bayes rule to compute the above equation.

$$\mu_1 \leftarrow \gamma\mu_1 + (1-\gamma)\frac{1}{n}\sum_{i|y_i=1} f_k\left(x_i\right) \quad (6)$$

$$\sigma_1 \leftarrow \gamma\sigma_1 + (1-\gamma)\sqrt{\frac{1}{n}\sum_{i|y_i=1}\left(f_k\left(x_i\right) - \mu_1\right)^2} \quad (7)$$

where γ is a parameter of learning rate. The update rules for μ_0 and σ_0 are defined in a way similar to that for μ_1, σ_1.