Computer Vision and Image Understanding 119 (2014) 41–56

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Co-trained generative and discriminative trackers with cascade particle filter ☆

Thang Ba Dinh [a,*,1], Qian Yu [b,1], Gérard Medioni [c]

[a] KLA-Tencor, 1 Technology Drive, Milpitas, CA 95035, USA
[b] SRI International, 201 Washington Road, Princeton, NJ 08540, USA
[c] University of Southern California, Institute for Robotics and Intelligent Systems Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 29 March 2012
Accepted 14 November 2013
Available online 8 December 2013

Keywords:
Visual tracking
Cascade particle filter
Co-training
Discriminative tracker
Generative tracker

ABSTRACT

Visual tracking is a challenging problem, as the appearance of an object may change due to viewpoint variations, illumination changes, and occlusion. It may also leave the field of view (FOV), then reappears. In order to track and reacquire an unknown object with limited labeling data, we propose to learn these changes online and incrementally build a model that encodes all appearance variations while tracking. To address this semi-supervised learning problem, we propose a co-training framework with cascade particle filter to label incoming data continuously and online update hybrid generative and discriminative models. Each of the layers in the cascade contains one or more either generative or discriminative appearance models. The cascade manner of organizing the particle filter enables the efficient evaluation of multiple appearance models with different computational costs; thus improves the speed of the tracker. The proposed online framework provides temporally local tracking that adapts to appearance changes. Moreover, it provides an object-specific detection ability that allows to reacquire an object after total occlusion. Extensive experiments demonstrate that under challenging situations, our method has strong reacquisition ability and robustness to distracters in clutter background. We also provide quantitative comparisons to other state of the art trackers.

1. Introduction

This paper aims at automatic visual tracking, *i.e.* once an object of interest is selected, our algorithm automatically tracks the object and reports a confidence which can be used to determine if the object is lost or out of field of view (FOV). When the object reappears, our algorithm reacquires the object and continues tracking.

We address three challenges in this problem:

1. Appearance Changes: Varying appearance, which can be caused by the changes in viewpoints, poses and illumination conditions, is one of the major challenges in visual tracking. New instances of the initially labeled object may constantly appear during tracking. Thus, visual tracking problem can be regarded as a weakly supervised learning problem. Very little supervised data is available in visual tracking. Improper updates of the appearance model (or no update) is the main reason of tracking drift, which is the most commonly seen failure in tracking.
2. Reacquisition: Persistent visual tracking requires the tracker to have the self-awareness of the status of tracking. A track is supposed to know if the object is out of FOV or is occluded, then reacquires the object when the object reappears. The solution requires an object-specific appearance model, in other words, a particular detector for "the" object, which has to be learned on-the-fly.
3. Time Performance: The success of visual tracking in recent years is mainly due to the powerful appearance models that have been used in visual tracking, such as [3,5,7,12,33]. However, the real-time performance of visual tracking is also an important factor in practice. A good balance of between the complexity of appearance models and the efficiency is desired.

We propose a co-training framework of generative and discriminative trackers with cascade particle filtering to address the above challenges. First, we formulate the appearance based object tracking as a semi-supervised learning problem: the process of selecting the object of interest before the automatic tracking can be considered as a process of providing labeled data in semi-supervised

☆ This paper has been recommended for acceptance by Carlo Colombo.

* Corresponding author.
E-mail addresses: thang.dinh@kla-tencor.com (T.B. Dinh), qian.yu@sri.com (Q. Yu), medioni@usc.edu (G. Medioni).

[1] This work has been done when the first and the second authors were studying for their Ph.D degrees at University of Southern California.

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.11.003

learning. Due to the appearance changes, the initially labeled data cannot fully represent the characteristics of entire distribution. A visual tracking approach needs to "learn to adapt" to the new appearance changes. Many visual tracking approaches are performed in a self-learning manner, where the sample with the most confident score evaluated by its own model is used to update itself. Here, we consider an example shown in Fig. 1 with a simple one dimensional distribution, where positive samples have two modes and negative samples contain one mode. All training samples are given sequentially. Except the few labeled training samples given at the very beginning, the rest of training samples are given as unlabeled. The dilemma of self-learning is shown in Fig. 1. If one adopts a strict threshold to update its model, the final model never learns new characteristics different from the initial labeled data. Thus, it will end up with a single mode either in zone a or zone c. On the other hand, if one adopts a loose threshold, its model is contaminated by outliers quickly and it will end up in zone b. Thus, self-learning is not a good way of weakly supervised online learning. Co-training proposed by Blum and Mitchell [18] is a principled semi-supervised training method. The basic idea is to train two classifiers on two conditionally independent views of the same data (with a small number of exemplars) and then use the prediction of each classifier to enlarge the training set of the other. It is proved that co-training can find an accurate decision boundary, starting from a small quantity of labeled data as long as the two feature sets are independent [18]. Empirical results [19] show that co-training also works well in the case where the independence is not perfectly satisfied. In our visual tracking setting, although our initial tracking samples are limited, if we regard multiple complementary features as approximated conditionally independent views of the same data, we can apply the co-training framework to combine multiple models to avoid the issues in the self-learning. One can certainly transform the semi-supervised learning problem to a supervised-learning problem for some specific applications: for instance, if the category of the object of interest is known, one can incorporate a model trained with a large amount of offline labeled data to compensate for limited online data, as in [30], or if tracking is allowed to be performed offline with human interaction, *eg.* directly adding new training data in a bootstrap manner as in [36]. These approaches go beyond the scope of general automatical visual tracking problem and require further information provided from user interaction.

Second, instead of combining multiple cues in a linear way, which increases the complexity linearly, we adopt cascade particle filter [30] to balance robustness and computational efficiency from multiple models. Instead of evaluating all models equally, this approach evaluates computationally cheaper models at earlier stages and more expensive models at later stages where much fewer particles remain. The cascade particle filter naturally combines with the co-training framework where multiple models need to be learned and evaluated on-the-fly. We call this proposed framework Co-trained Cascade Particle Filter (CCPF). Compared with co-training all features at the same stage, CCPF benefits from the robustness in co-trained multiple models and reduces the computational costs of different models. The CCPF framework is shown in Fig. 2.

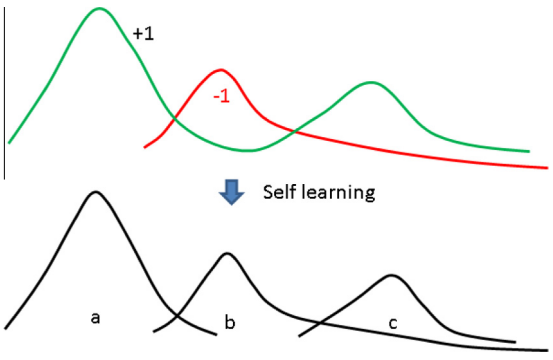

Fig. 1. Issues with self-learning.

Third, while the CCPF framework separates various features into different stages, the last stage of the CCPF makes the final decision for object reacquisition. Thus, besides the tracking capability, the end-product of the tracker is also a detector of the particular object that has been tracked. The detector contains all the appearance variations of the object that have been observed since tracking is started, and can be used to reacquire the object once it reappears.

The rest of this paper is organized as follows. The related work is presented in Section 2. The overview and the advantages of our proposed framework are presented in Section 3. All of the online appearance models of trackers are described in Section 4. Then the experiments are shown in Section 5, followed by summary and future work.

2. Related work

Both generative and discriminative learning approaches have been extensively used in visual tracking. Several examples of generative tracking algorithms are Eigentracking [1], WSL tracking [2], Incremental Visual Tracking (IVT) [3], and L1 Minimization Tracking [40]. Due to the fact that appearance variations are highly nonlinear, multiple subspaces [4] and non-linear manifold learning methods [5] have been proposed. Due to background information is too extensive to represent in a generative model, most traditional generative tracking methods are merely trained based on object appearance without considering background information. However, generative approaches are capable of dealing with partial missing data. In the visual tracking problem, missing data occurs when an object is occluded.

Instead of building a generative model to describe an object itself, discriminative tracking methods aim to find a decision boundary that can best separate the object from the background. Recently, many discriminative trackers have been proposed [7–9,33] and demonstrate strong robustness in avoiding distracters in the background. Support Vector Tracking (SVT) [6] integrates an offline trained Support Vector Machine (SVM) classifier into an optic-flow-based tracker. In order to update the decision boundary according to new samples and background, discriminative tracking methods with online learning have been proposed in [8,7]. In [8], a confidence map is built by finding the most discriminative RGB color combination in each frame. However, a limited color feature pool restricts the discriminative power of this method. In [7], Avidan proposed an ensemble of online learned weak classifiers to label a pixel as belonging to either the object or the background. To accommodate for object appearance changes, in every frame, new weak classifiers replace part of old ones that do not perform well, or have existed longer than a fixed number of frames. Both methods [8,7] use features at the pixel level and rely on a mode seeking process (mean shift) to find the best estimate on a confidence map, which restricts the reacquisition ability of these methods. Oza and Russell [10] proposed an online boosting algorithm, which is applied in the visual tracking problem [11,12]. Due to the large number of features, either an offline feature selection procedure or an offline trained seed classifier is required in practice. Thus, it is difficult to generalize to arbitrary object types using tracking methods based on online boosting. More recently, Tang et al. [20] proposes to use co-training to online train two discriminative trackers with color histogram features and HOG features. It has been shown that discriminative classifiers out-