Pattern Recognition 69 (2017) 169–183

Contents lists available at ScienceDirect

Pattern Recognition

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

Dynamic multi-level appearance models and adaptive clustered decision trees for single target tracking

Jingjing Xiao[a,b,*], Rustam Stolkin[b], Aleš Leonardis[b]

[a] Department of Medical Engineering, Xinqiao Hospital, Third Military Medical University, Chongqing 400037, China
[b] University of Birmingham, B15 2TT, U.K.

ARTICLE INFO

Article history:
Received 9 October 2016
Revised 14 February 2017
Accepted 4 April 2017
Available online 14 April 2017

Keywords:
Single target tracking
Adaptive clustered decision trees
Multi-level appearance models

ABSTRACT

This paper presents a tracking algorithm for arbitrary objects in challenging video sequences. Targets are modelled at three different levels of granularity (pixel, parts and bounding box levels), which are cross-constrained to enable robust model relearning. The main contribution is an adaptive clustered decision tree method which dynamically selects the minimum combination of features necessary to sufficiently represent each target part at each frame, thereby providing robustness with computational efficiency. The adaptive clustered decision tree is used in two separate ways: firstly for parts level matching between successive frames; secondly to select the best candidate image regions for learning new parts of the target. We test the tracker using two different tracking benchmarks (VOT2013-2014 and CVPR2013 tracking challenges), based on two different test methodologies, and show it to be more robust than the state-of-the-art methods from both of those tracking challenges, while also offering competitive tracking precision. Additionally, we evaluate the contribution of each key component of the tracker to overall performance; test the sensitivity of the tracker under different initialization conditions; investigate the effect of using features in different orders within the decision trees; illustrate the flexibility of the method for handling arbitrary kinds of features, by showing how it easily extends to handle RGB-D data.

1. Introduction

After several decades of visual tracking research, even the most sophisticated trackers are still prone to failure in challenging scenarios, including clutter and camouflage in one or more feature modalities, rapid and erratic target motion, occlusions, and targets which change their shape and appearance over time. These problematic tracking conditions predominantly lead to failures in three fundamental parts of the tracking algorithm: 1) the representation or model of the target object's visual appearance; 2) the mechanism for matching model parts to image regions at each frame; 3) the mechanism for continuously relearning or updating models of targets which change their appearance over time.

This paper presents a target tracking algorithm which achieves state-of-the-art robustness by addressing each of these three fundamental areas. We propose a flexible target representation which can adaptively exploit an arbitrary number of different image features. Targets are modelled at three different levels of granularity, including the level of individual pixels, the level of local parts (constructed from super-pixels), and a bounding box level which encodes overall information about the target as a whole. Cross-constraints between these different levels during updates enable continuous target model relearning which is robust and stable.

The main contribution of the paper is an adaptive clustered decision tree approach which dynamically selects the minimum combination of features necessary to sufficiently represent each target part at each frame, thereby providing robustness without sacrificing computational efficiency. We show how this adaptive clustered decision tree can be utilised in two separate parts of the tracking algorithm: firstly to enable robust matching at the part level between successive frames; and secondly to select the best candidates (constructed from super-pixels) for learning new parts of the target. During *matching*, the adaptive clustered decision trees are used to search the set of candidates in the current frame, to find the best match to a target part in the previous frame. During *model updating*, the decision trees are used to search for the most suitable candidate to model a new part of the target, and to replace an old target part which drifts from the main body of the target in the current frame.

We have carried out a principled evaluation using the latest benchmark methods, and comparing against the other state-of-the-art trackers. Results show that the proposed method outperforms the best trackers on both VOT2013 and VOT2014 benchmark sets.

[*] Corresponding author.
E-mail address: shine636363@sina.com (J. Xiao).

http://dx.doi.org/10.1016/j.patcog.2017.04.001
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

It outperforms the 7 available methods on the CVPR2013 dataset w.r.t. robustness, while also achieving competitive tracking accuracy. Furthermore, we have decomposed the tracker to evaluate the effectiveness of each component, and evaluated the tracking performance with the various noisy initialization conditions. To have a deeper understanding of the proposed adaptive clustered decision trees, we also implemented the tracker on the publicly available RGB-D sequences and showed that, with well designed clustering methods, the tracker is relatively robust within various feature sequential orders.

The rest of the paper is organized as follows. Related work is discussed in Section 2. The multi-level target model, and its initialisation, is introduced in Section 3. Section 4 explains the adaptive clustered decision tree, and shows how it is used for both target matching and model updating at each successive frame. Section 5 presents and discusses the experimental results of testing the tracker on the VOT2013, VOT2014 and CVPR2013 benchmark video datasets. Additional experiments analyse the contributions of each key part of the tracker, to help explain the strong overall performance. Further more, we investigate how the noisy initialization affects the tracking performance. Section 6 shows how the decision trees can easily be extended to include arbitrary kinds of additional features, e.g., depth feature, illustrating how the decision trees dynamically vary the number of features exploited for each target part at each video frame. We further investigate the extent to which the tracker is robust to the order in which different features are represented at different tree levels. Section 7 provides concluding remarks and mentions ongoing efforts to extend this work in various new directions.

2. Related work

In this section, we review recent tracking algorithms in terms of three primary components: target representation, matching mechanism, and model update mechanism. We also discuss the relationship between our proposed adaptive clustered decision tree method and other kinds of tree-like, hierarchical or recursive classification methods from the literature.

2.1. Target representation

Choice of target representation is a crucial component of any tracker. Two main streams of research can be distinguished. The first uses holistic (overall) target templates for tracking, e.g., [27,40]. However, such methods have difficulty in handling significant appearance changes and deformations of the target. Later work [18,20], proposed patch-based approaches to provide more flexibility for target matching. However, the choice of geometric constraint for the local patches' movement remains an open problem, while environmental clutter can often distract such local patches and cause them to drift. [38] avoided complex geometric constraints for patch motion, by treating the problem as a classification of foreground and background super-pixels. However, since each super-pixel is classified independently, this method remains prone to failure with cluttered background scenes. In contrast, our method also makes use of super-pixels, but exploits them within a more robust cross-constrained multi-level target model structure. More recent work [35] also combined both holistic and patch-based target models together for a more robust representation. However, this work fused multiple features in a homogeneous way (i.e., equally weighting the opinions of all feature modalities), which causes failures under conditions where one or more features become less discriminating than others. In contrast, our method achieves better results by adaptively selecting in favour of whichever feature or feature combination is most discriminating for each target part in each new frame.

2.2. Matching

To estimate the state of the target, the algorithm must match observations from a candidate image region to the target representation model. A single feature is not sufficient to handle large appearance variations, and recent work [23,29,33], increasingly exploits combinations of multiple features. One approach is to compute the likelihood from all features and then multiply all values to estimate the target state [35]. However, in such schemes, a poorly performing feature can degrade tracking performance, even when other features are highly discriminative. Therefore, instead of treating all features with equal importance, other methods, e.g. [8,33,37], attempt to identify and weight in favour of the most discriminative features at each time step. Brasnett et al. [6], propose a scheme for weighting in favour of the best performing features, and updating these weights adaptively at each new frame. However, this method ignores feature saliency from the local background regions. In contrast, recent work [33] proposes an adaptive method which successfully exploits contextual information for optimally weighting the contributions from each feature online during tracking. However, [33] uses only a simple holistic target model which is insufficient to cope with large target deformations and appearance changes. Pernici et al. [29] propose a matching method which uses both the target and context SIFT features. However, the matching indices are obtained directly by a nearest neighbour search, which might perform poorly when the target undergoes rapid and significant deformations and appearance changes. In contrast to the homogeneous treatment of all features by e.g. [35], our adaptive clustered decision tree approach can adaptively select in favour of the most discriminative features for matching each target part to each new frame. Moreover, this adaptive feature selection is also embedded within a cross-constrained multi-level target representation, which enables much more robust matching and model updating than the simple holistic target models of e.g. [33].

2.3. Model update mechanisms

For robust, general tracking, it is essential to continuously update or relearn the target model to cope with appearance changes. An appropriate target model should enable the tracker to overcome errors in the relearning process which might corrupt the target model, and support long term tracking without drifting [24]. Early methods, such as [27], updated the model at every frame as a simple linear combination of the previous model and the most recent estimation of the target region in the current image. Without additional methods for precise delineation of the target parts, such update methods are likely to fail, given sufficiently long tracking duration, due to accumulated errors and noise during successive updates. In MIL [4], and other trackers such as OB [12] and SB [13], updating of the target model is performed by an evolving boosting classifier that tracks image patches and learns the object appearance. However, online boosting requires that the data should be independent and identically distributed, which is a condition not satisfied in most real video sequences, where data is often temporally correlated [29]. A more robust updating mechanism is achieved by [35], which forms a cross-constraint paradigm to stably constrain the relearning of a two-layer target model, in which global (bounding region) and local (parts based) models are used to constrain (and thereby stabilise) each others' online relearning. However, this method (and most earlier methods e.g. [27]) updates target appearance models at a fixed rate, regardless of the confidence (or lack of) in current target observations. This problem is compounded by the previously described problem, that many methods, e.g. [35], combine the opinion of all features with equal weight, which can lead to drifting of patches away from the true target