Signal Processing: Image Communication 68 (2018) 169–180

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Weak-structure-aware visual object tracking with bottom-up and top-down context exploration

Ning Liu [a], Chang Liu [a], Hefeng Wu [b,*], Hengzheng Zhu [a], Jin Zhan [c]

[a] Sun Yat-sen University, Guangzhou 510006, China
[b] Guangdong University of Foreign Studies, Guangzhou 510006, China
[c] Guangdong Polytechnic Normal University, Guangzhou 510665, China

ARTICLE INFO

Keywords:
Visual tracking
Weak structure
Keypoint
Context exploration

ABSTRACT

It poses great challenges to model-free trackers that the object undergoes large appearance variations due to motion, shape deformation, occlusion and surrounding environments. In this paper, we investigate a novel method for modeling and locating the object by being aware of the weak structures of discriminative parts of both the object and its surroundings. The discriminative parts are modeled based on keypoints and feature descriptors. We separate the discriminative parts into two sets corresponding to object and background, and model their spatial structure relationship with the object. While tracking, the successfully localized parts will contribute to potential centers of the object. Aware of the weak structures, we further cluster potential centers to locate the object. The object scale is also updated adaptively. To increase the accuracy of this weak-structure-aware location inference, we fully explore context in both bottom-up and top-down procedures. In the bottom-up stage, we explore the local motion estimation of low-level pixels. The bottom-up information produces consistent tracking of discriminative parts. In the top-down stage, we build a superpixel kernel model to roughly distinguish the object from its surroundings, which provides guided information for location inference and model update. The effectiveness of the proposed method is verified by evaluation on a popular benchmark and comparison with recent tracking methods.

1. Introduction

Visual object tracking is a fundamental task in many high-level applications such as intelligent monitoring [1], automatic driving [2] and human–computer interaction [3]. Although significance progress has been witnessed in recent years, it remains challenging when the tracked object undergoes large appearance variations in complex circumstances. The task becomes harder still for a model-free tracker [4,5] aiming to track an object of arbitrary class.

Recently, part-based object appearance models have attracted considerable attention because of their flexibility in tackling large appearance variations. Part-based models aim to overcome the limitations of holistic ones by considering the tracked object as an ensemble of several parts, in which the localization of the other parts can still be effective even when appearance variations render certain parts useless.

However, two key issues need to be addressed by a part-based tracker: (i) how the parts of the object are modeled and tracked, and (ii) how the tracked parts are combined to localize the holistic object. In model-free tracking, the tracked object is commonly given by a rectangular bounding box in the first frame of a video sequence, rendering these problems more challenging. First, little information is available about the parts and structure of the object. Second, the object's appearance changes over time are unpredictable.

To address the aforementioned issues, many part-based tracking solutions have been proposed [6–9]. The target object is commonly represented by a set of overlapping or nonoverlapping rectangular patches [6,7], keypoints [8], blobs [10], etc. Most previous works assume strong structures of the object's parts (i.e., rigid geometric structure relationship), and the object is located by weighting the average of the parts under the structure assumption [6] or minimizing the model transform error (e.g., with RANSAC estimation) [8]. However, this assumption is often not valid in challenging situations such as large pose changes, out-of-plane motion, and non-rigid deformation. Moreover, the object localization under the strong structure assumption is error-prone due to drift-away parts that are wrongly considered successfully tracked or background parts that are incorrectly incorporated to represent the

* Corresponding author.
E-mail address: wuhefeng@gmail.com (H. Wu).

https://doi.org/10.1016/j.image.2018.08.005
Received 22 December 2017; Received in revised form 9 August 2018; Accepted 9 August 2018
Available online 13 August 2018

object. These problems remain difficult in the literature of part-based trackers.

Therefore, in this paper, we investigate a novel part-based object tracking method and provide an effective solution based on a weak-structure-aware strategy of discriminative parts. We model the discriminative parts based on keypoints and feature descriptors, which are a feature vector extracted in a local image patch and have recently been shown to have remarkable discriminative ability [11–13]. These descriptors have great characteristics (e.g., invariance to scale, rotation and illumination) that are quite useful in tackling the challenges faced in object tracking. We build the part-based model with two set of discriminative parts, one for the object and the other for the surroundings. While the part set of the object aims to keep the tracking effective and stable, the part set of the surroundings works in a complementary way and is especially helpful under occlusion. We introduce a weak-structure-aware strategy that assumes no rigid structure relation for the parts. This strategy can automatically discover stable parts for object localization. Specifically, the parts considered successfully tracked produce potential object centers and the object location is found via clustering and weighting. The scale of the object is also adaptively updated.

Furthermore, we devise bottom-up and top-down context exploration procedures to couple with the weak-structure-aware tracking. The localization of discriminative parts is achieved by keypoint matching. However, this step can be degraded by such challenging situations as fast motion or sudden large variations. Thus, we introduce a bottom-up context exploration stage, which estimates the motion of local low-level pixels to facilitate consistent tracking of the parts. On the other hand, to maintain the two part sets of object and surroundings, we introduce a top-down context exploration stage and propose to build a superpixel kernel model at the object level to roughly distinguish the object from background context.

In summary, this work has the following contributions. First, we introduce a weak-structure-aware solution to part-based object tracking, which we call the Weak-Structure-Aware Tracker (WSAT). Second, we model the part-based representation with two sets of discriminative parts, corresponding to the object and its surroundings respectively, which are combined complementarily in a weak-structure-aware scheme to find the object location and scale adaptively. Third, we employ bottom-up and top-down procedures to explore pixel-level and object-level context information for effective supports. Moreover, the proposed WSAT tracker is verified in extensive experiments to demonstrate its effectiveness.

The rest of the paper is organized as follows. We review the related work in Section 2. Afterwards, Section 3 describes the proposed method in detail, and experiments are given in Section 4. We conclude the paper and discuss future work in Section 5.

2. Related work

We will briefly review the most relevant literature of model-free object tracking in this section. Some excellent recent surveys [14,15] and benchmark datasets (e.g., OTB [16], VOT [17], NUS-PRO [18]) are recommended to interested readers. With respect to holistic appearance modeling, the existing tracking methods generally represent the object by generative or discriminative models.

Generative models [19–22] focus on a good representation of the tracked object and track the object by finding the most similar candidate to the built model, e.g., kernel-based histograms [19] and Gaussian mixture models [20]. However, these methods may easily drift away when the object model is not discriminative enough from the surroundings. Recently, sparse representation [23–26] has been introduced into object tracking and has achieved favorable results, though limited by the construction of over-complete dictionaries and the high computational load of $\mathcal{L}_1$ minimization.

By contrast, discriminative models [27–31] try to distinguish the tracked object from the background, which can be formulated as a binary classification task. This category of tracking methods have made promising progress by taking advantage of recent advances in machine learning theory, e.g., online AdaBoost [27], multiple instance learning [28] and structured output support vector machine (SVM) [29]. It is worth mentioning that recent correlation filter-based tracking methods [31–33] achieve high efficiency by transferring to the frequency domain for model training and object localization. Another impressive progress is employing deep neural networks with pre-trained models, which can achieve state-of-the-art tracking performance due to their strong feature learning ability [34–38]. However, in model-free object tracking, the problem of lacking good training samples remains challenging for these methods.

In this paper, we investigate part-based models for object tracking. Part-based models [6,7,39,40] are an attractive alternative to holistic object models, and as a result they have been receiving increasing amounts of research attention in recent years. As previously mentioned, part-based approaches model the object as a set of separate parts, motivated by the key assumption that not all of these parts will be undergoing large movements/deformations at the same time. By modeling each part using a generative or discriminative model, and using parts that are being successfully tracked to guide the tracking of more difficult parts, they aim to maintain successful tracking of the object as a whole even when the tracking of some individual parts is failing. In [41], a fast adaptive algorithm is proposed for tracking non-rigid objects, and this method is quite novel for the use of the generalized Hough transform with pixel-based descriptors. Adam et al. [6] divide the bounding box containing the object with a pre-defined grid and each cell of the grid represents a part. This straightforward method demonstrates the effectiveness of part-based methods. However, the fixed spatial layout makes it difficult for the model to tackle certain types of appearance deformation. By using the fast correlation filter tracker [31] to track each part, the part-based tracker presented in [40] adopts a Bayesian inference framework and a structural constraint mask to combine the parts and handle various appearance changes. However, they assume a fixed number of part models at initialization, which makes it difficult to handle large appearance changes of object. In contrast to these methods, we model the part-based representation with a flexible number of discriminative parts in this paper and combine them with a weak-structure-aware strategy to locate the object.

With the good feature descriptors that can describe local patches well (e.g., SIFT [11], SURF [12] and ORB [13]), many tracking methods utilize keypoints for object tracking. In [8], Hare et al. fuse the keypoint matching and geometric transformation estimation into a single structured output learning framework. Pernici and Del Bimbo [42] use multiple instances of keypoints that weakly align along the object template to build discriminative classifiers for the object modeling. These methods can track rigid objects robustly, but they are not suitable for tracking non-rigid objects. Nebehay et al. [43] construct an object tracker based on keypoint voting. The object model in this tracker is initialized in the first frame to learn the spatial relations between keypoints and object location. However, they do not update the keypoints during the tracking, which may make it difficult for the object model to adapt to appearance changes and irregular deformations.

3. Proposed method

This section describes the details of the proposed method. Our method takes a sequence of consecutive image frames I_t $(t = 1, 2, \ldots)$ as input. Before tracking, the object is marked by a rectangular bounding box in the first frame manually or by an object detector for initialization. We use middle level image features (i.e., keypoints and associated feature descriptors) to build the part-based model for object tracking. We maintain two sets of keypoints, P_O for object and P_B for background. By matching keypoints between two adjacent frames, we can predict the trend of each parts of object, and then locate the object using the structure information of the object and the background in a