Neurocomputing 202 (2016) 1–11

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-level visual tracking with hierarchical tree structural constraint

Jingjing Wang, Nenghai Yu, Feng Zhu, Liansheng Zhuang*

CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei, China

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form
4 March 2016
Accepted 10 March 2016
Available online 26 March 2016

Keywords:
Object tracking
Part-based model
Multi-level representation
Hierarchical tree
Structural constraint

ABSTRACT

Recently, part-based model has drawn much attention in visual tracking for its promising results in handling occlusion and deformation. However how to divide the target into parts and how to model the relationships between parts are still open problems. In this paper, we propose a robust tracker based on multi-level target representation and hierarchical tree structural constraint. The multi-level target representation models the target at three different levels: the bounding box (top) level, the superpixel (middle) level and the keypoint (bottom) level. The relationships between parts at all levels are modeled by the proposed hierarchical tree which includes intra-layer and inter-layer structural constraints. The positions of all the parts are optimized jointly in a unified objective function taking into account both the appearance similarity and the hierarchical tree structural constraint. The appearance model and the hierarchical tree structure are updated online to adapt to the changes of the target in both appearance and structure. Extensive experiments on various challenging video sequences demonstrate that the proposed method outperforms the state-of-the-art trackers significantly.

1. Introduction

Visual tracking plays a critical role in a wide range of applications in computer vision, such as surveillance, human computer interactions and robotics. Although visual tracking has been studied for decades, it is still a challenging task to track an arbitrary object in a given video sequence due to many challenging issues, including occlusion, deformation, illumination variation, scale and pose change and background clutter. Among the challenging issues, occlusion and deformation have drawn much attention, because they are ubiquitous in real world tracking scenarios and remain unsolved [1]. Recently, an increasing number of tracking algorithms are proposed to solve the occlusion and deformation problems, of which part-based model is mostly adopted by state-of-the-art trackers [1–9].

For part-based model, how to divide the target into parts is critical. Different strategies have been proposed to generate local parts. According to the level of granularity, part-based target representation can be categorized into three levels: top level (holistic target models [10–12]), middle level (superpixels [1,5,3] and patches [13,2,6,9]), and bottom level (keypoints [7,14,15] and pixels [16,17]). Each level target representation has its pros and cons. For example, lower-level representation based trackers can handle occlusion and deformation better, but perform relatively poorly in scenarios where there is excessive background clutter due to the lack of holistic appearance of the target. In contrast, higher-level representation based trackers are robust to background clutter and partial occlusion, but tend to fail when the target undergoes severe occlusion and non-rigid deformation. In a word, only a single-level target representation cannot be suitable for all objects in all scenarios [4]. However, only a few methods [4,8,18] utilize multi-level target representation.

For general objects, although the appearance may change drastically over time due to the multiple challenges, the structure relationships inside the target remain stable [19]. So incorporating geometric constraints is very helpful to enhance the stability of a tracker. However, many methods [3,6,9] ignore the structure relationships between local parts and treat the parts independently. These methods are prone to failure with cluttered background. Recently, Cai et al. [1] use an undirect graph to model the relationships between superpixels. Xie et al. [19] represent the structure relationships between superpixels by a minimum spanning tree. However, these two methods use only a single-layer representation, and the inter-layer structures with multi-layer representation are unexplored. To this end, Yao et al. [18] model the relationships between the holistic object template and a fixed number of local patches using a star model, while the dependencies between local parts are ignored. In challenging scenarios, fully utilizing multi-layer constraints including intra-layer and inter-layer ones can make the tracker more robust than only using one type of them. However, multi-layer structural constraints with multi-layer representation are not well explored.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.03.010
0925-2312/© 2016 Elsevier B.V. All rights reserved.

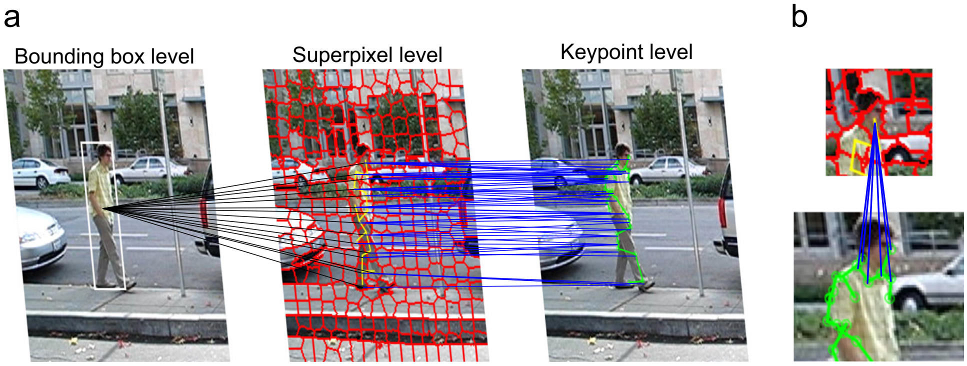

Fig. 1. (a) shows the illustration of the multi-level object representation and hierarchical tree structure. The target is represented at three levels: bounding box level, superpixel level and keypoint level. At the bounding box level, the target is represented by a bounding box (indicated by a white rectangle). At the superpixel level, the target is represented by superpixels inside the bounding box. At the keypoint level, the target is represented by keypoints inside the bounding box (indicated by green circles). Different levels are constrained by a hierarchical tree including intra-layer structures and inter-layer structures. Intra-layer structures: the relationships between neighboring parts at both superpixel level (indicated by yellow lines) and keypoint level (indicated by green lines) are modeled by a minimum spanning tree; inter-layer structures: the relationships between the bounding box and the superpixels (indicated by black lines) and the relationships between superpixels and their nearby keypoints (indicated by blue lines) are modeled by star models. To make the connections between superpixels and keypoints more clear, (b) shows a superpixel corresponding to the face region of the target is connected with the nearby keypoints by a start model. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Motivated by the above observations, we propose a novel robust part-based tracker, which represents the target at multiple levels and models the structure relationships between parts by constructing a hierarchical tree (Fig. 1). Specifically, the target is divided into parts at three levels: the top level (bounding box), middle level (superpixels) and bottom level (keypoints). At the top level, the target is represented by target templates to capture the holistic information of the target; at the middle level, the target is segmented into many superpixels as the middle level parts; at the bottom level, keypoints are extracted inside the target to serve as the smallest parts. In the hierarchial tree, the structure relationships between neighboring parts at the superpixel level and the keypoint level are modeled by a minimum spanning tree (MST). Beyond these intra-layer structures, the inter-layer structures are employed to make the structural constraints more stable. The target bounding box is connected with all the superpixels by a start model, and the superpixels are also connected with their nearby keypoints by start models. In this way, we fully utilize the target's parts at different levels and structural constraints at and between different levels to represent the target. After constructing the robust target representation, we can infer the target state via voting from these parts. The optimal locations of the parts at all levels are obtained by quantifying the appearance similarity and the deformation cost, and optimized jointly in a unified objective function. Thus, each level can benefit from other levels, and the overall performance of the tracker is enhanced. During tracking, the target may undergo various appearance changes. To adapt to the appearance changes, the appearance model and the hierarchical tree structure are updated online.

In summary, the main contributions of this paper include: (1) We propose a novel multi-level target representation which represents the target at three different levels and fully utilizes the advantages of all levels. (2) We propose a hierarchical tree structure to quantify the geometry structure relations of parts. The hierarchical tree bridges the multi-level parts through intra-layer structural constraints and inter-layer structural constraints to help track individual parts more stably. (3) The locations of different level parts are optimized jointly in a unified objective function taking into account the appearance similarity and the hierarchical tree structural constraint. (4) The appearance model and the hierarchical tree structure are updated online to adapt to the appearance and structure changes of the target.

The reminder of the paper is organized as follows. Section 2 reviews the related works. Section 3 describes our proposed method including multi-level object representation, hierarchical tree structure, object state inference and update strategy. Experimental results are presented in Section 4, and some conclusions in Section 5.

2. Related work

In visual tracking, target representation is very important. First, some features are extracted to represent the target, such as intensity [10], color [20], and Haar-like features [21]. To make the target representation more efficient and effective, the dimensionality of the feature can be reduced by feature selection methods [22,23] such as in [24,25], or Hashing methods [26,27] as in [28–30]. Then, given the extracted features, the target can be represented holistically or locally. Some recent reviews of visual tracking can be found in [31,32].

Most previous tracking methods represent the target using holistic templates at the highest level [10–12,33]. Ross et al. [10] propose to learn an incremental subspace model to represent the target, which is robust against the illumination variations. However this method is less effective for handling occlusion as a result of the holistic appearance model. To enhance the robustness against occlusion, Mei et al. [11] use the sparse representation to reconstruct the target from a template set, in which trivial templates are added in order to handle partial occlusion. The trivial templates make the computation cost of the method in [11] very high. Moreover, the trivial templates can model not only the target but also the background. To solve these problems, Zhang et al. [33] propose to learn a variation dictionary online instead of trivial templates for visual tracking. In [21], a multiple instance learning strategy is adopted to minimize the effects of occlusion by collecting bags of image patches as the training samples. Other methods [34–36] exploit the context information to overcome the occlusion problem. However, none of the above trackers consider deformation.

Compared to rigid templates, local parts are more flexible and less sensitive to deformation and occlusion. Recently part-based model has attracted much attention and has shown promising results for handling deformation and occlusion.

At the lowest level, pixels and keypoints are commonly used to represent the target. Avidan [16] formulates the target tracking problem as a pixel-based classification problem. An ensemble of weak classifiers is trained online to classify individual pixels as the