Signal Processing: *Image Communication* 36 (2015) 43–52

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Geometric structure-constraint tracking with confident parts

Zhao Xie*, Jing Chen, Tingting Yao, Yongxuan Sun

Laboratory of Image Information Progressing, School of Computer and Information, Hefei University of Technology, No.193, Tunxi Road, Hefei, Anhui Province 230009, China

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form
3 June 2015
Accepted 3 June 2015
Available online 11 June 2015

Keywords:
Visual tracking
Structure representation
Dynamic programming
State estimation

ABSTRACT

Recent work has successfully tracked objects with single challenge by matching similar features or classifying positive samples. However, if appearance of object changes drastically when it simultaneously encounters multiple challenges especially occlusion, fast motion and deformation in real-world, it is hard to track object via similar appearance representation. Instead of finding suitable match, we propose an adaptable structure representation tracking (SRT) method to infer object state by exploiting confident parts of object based on mid-level appearance and structure constraints. The structure representation is graph-based and propagated by dynamic programming. The experimental results of our SRT demonstrate significant tracking performance especially in challenging environments.

1. Introduction

Visual object tracking is one of the fundamental challenges in computer vision with wide-ranging applications including scene surveillance, video retrieval, human–computer interaction, and robot navigation. Although significant progress has been made in tracking specific objects, tracking generic objects in real-world remains hard. The main problem is that appearance variance is caused not only by fast motion or abrupt deformation of objects but also by external influences such as illuminations, part or full occlusions and changes of camera viewpoint. The changes of appearance make features change dramatically, therefore it makes trackers, which find object via feature similarity, fail the tracking [1].

In order to achieve robust tracking, selecting appropriate features to capture appearance changes of object is important. Generally speaking, the representation for object in video sequences can be classified into three levels: low-level, mid-level and high-level. Low-level appearance description directly employs original pixel information or codes intensity and color information. For example, some methods are based on a single feature such as intensity [2,3] and gradient [4] of pixels. Others try to synthesize multiple complementary low-level features [5] to improve performance. It works for accurate instance matching [6] but does not suit real scene tracking because intensity of each pixel changes drastically even though the object only changes slightly. Mid-level object representation exploits image patches or parts obtained by clustering similar pixels as basic unit or histograms quantification to merge image blocks and it makes the model representation robust to object changes [7]. High-level methods consider semantic meaning for samples, but it lacks adaptive capacity and discriminative ability especially when object is blur or occluded. Therefore it is prone to drift [8–10] and that makes it hard to apply for tracking fast motion and complicated deformation.

As single level is unlikely to suit for tracking all the objects, Ref. [11] exploits multiple levels to quantify appearance of object and finds the most possible position of the target by jointly classifying the pixels and superpixels and obtaining the best configuration across all levels. Compared with sole pixel-level representation, the

* Corresponding author. Fax: +86 551 62901393.
E-mail address: xiezhao@hfut.edu.cn (Z. Xie).

http://dx.doi.org/10.1016/j.image.2015.06.001
0923-5965/

employment of mid-level representation makes the model applicable for generic object and robust to changes such as object motion, lighting conditions and occlusion in running, so we select mid-level feature as basic representation of object in this paper.

People can locate object by merely seeing partial object and it is benefited from the ability to infer object state with prior knowledge of structure relationship between parts within object. For generic object especially non-rigid one, when it deforms, the majority of the adjacent parts are still close to each other. For example, when a man walks or jumps, his head preserves on the torso and legs are below torso though the pose of legs may change drastically. By means of tree structure, relative movement between parts of object can be flexibly modeled and the extent of deformation of object itself can also be measured. Moreover, when appearance of partial object is influenced by illumination, we can infer position of this part via neighboring parts. As we focus on the relationship of parts with their neighbors which are not only adjacent to each other but also similar to each other on appearance, we employ Minimum Spanning Tree (MST), one type of graphs, to represent the geometric structure constraint in practice.

After constructing robust representation for object, we infer object state via confident parts which are obtained by quantifying the appearance similarity and extent of deformation. Gained the inferred object state, we employ update scheme to evaluate the correctness of inference and judge occlusion. We handle occlusion in two aspects. Firstly, we judge whether some parts of object are occluded and we abandon the contribution of these parts in inference of object state. In realistic situation we can know exact position of a car when we merely see the front part of a car and the other parts are occluded by buildings. So even object is occluded, we can employ the parts which are not occluded to infer object state. Secondly, after obtaining object state via inference, we will determine whether to update object model to adapt to appearance change according to the correctness of inference. In these two ways, we can estimate the occurrence of occlusion and do not update the heavily occluded object because appearance of the object which includes a lot of background information will disturb the following tracking.

The main contribution of this paper is as follows: First, based on mid-level representation, we employ MST to quantify geometric structure relation of parts in object therefore our tracking method can measure deformation extent and improve tracking. Second, by means of dynamic programming, we integrate the appearance change of parts following structure relation to evaluate the extent of certainty of tracked object instead of feature similarity. Third, we merely extract confident parts instead of the entire parts of object to infer object state and it is beneficial for eliminating the influence of background in tracking window. Moreover, our algorithm framework can judge occlusion and decrease its influence for tracking objects and updating model.

The rest of this paper is organized as follows: Section 2 discusses the related works. Section 3 describes structured representation model, object state inference and update scheme in detail. We present the experimental results in Section 4 and conclude this paper in Section 5.

2. Related works

There are great research efforts on tracking algorithms, especially focusing on the specific challenges including occlusion [12], deformation [4,13] and real-time processing [14,15] to design tracking algorithms. Most of the popular approaches treat tracking problem as a classification task and employ online learning techniques to update object model, popularized by [14] and [16]. Large improvements benefit from optimizing learning algorithm to discover and correct the wrong example classification labels such as [17]. Moreover, robust features for object representation are also important for distinguishing object in feature space. We refer to detailed survey on [1,18].

For faults of low-level and high-level features mentioned in Section 1, mid-level representations have attracted attention in recent researches. Ref. [12] segments the object within rectangle into irregularly shaped superpixels and judges occlusion via the confidence map of superpixels. Ref. [19] uses patches sampled from object as local information to reconstruct histogram representation and tackles occlusion influence via refusing patches which have reconstruction error. Ref. [20] exploits dynamic graph to represent neighboring relations and thereby obtains histograms of each superpixel and its neighboring superpixels. In this way they optimize the similarity of histograms caused by adding or subtracting over-segmented superpixels to obtain object. Ref. [21] averagely divides tracking window into four regions and employs superpixel to obtain color and boundary features of each region, then they infer the position of object via these feature similarity. Ref. [13] employs relational hypergraph to model appearance of local parts in multiple consecutive frames. Also they use the SVM-based Part Classifier to preserve historical appearance information and identify local parts which are excluded by dense neighborhoods searching. When object is partially occluded, tracker of [13] will shrink to the non-occluded parts and continue tracking target after occlusion with the help of historical appearance information. Although these trackers mentioned above can improve tracking performance by focusing on stable similar representation, when appearance of object changes dramatically caused by multiple challenges simultaneously, the representation of the identical object is not similar enough anymore.

Therefore we try to find confident parts which preserve stability to a certain extent even encounter deformation, illumination or partial occlusion to infer object state. Although researches mentioned above consider the influence of parts to object, they merely focus on local appearance of all parts and neglect relation of spatial structure between parts. For generic objects, structure relationship inside object preserve stability though appearance changes drastically over time. Ref. [4] tracks multiple objects by employing spatial relation between multiple objects to promote inference of each single object. Inspired by [4], for tracking single object, we consider structure relationship between parts of object to