Pattern Recognition 46 (2013) 2187–2201

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Visual tracking by proto-objects

Zhidong Li [a,b], Weihong Wang [a,b], Yang Wang [a,b,*], Fang Chen [a,b], Yi Wang [a,b]

[a] National ICT Australia[1], Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia
[b] The University of New South Wales, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 10 April 2012
Received in revised form
31 October 2012
Accepted 14 January 2013
Available online 26 January 2013

Keywords:
Tracking
Proto-object
Saliency
Gibbs sampling
Bayesian

ABSTRACT

In this paper, we propose a biologically inspired framework of visual tracking based on proto-objects. Given an image sequence, proto-objects are first detected by combining saliency map and topic model. Then the target is tracked based on spatial and saliency information of the proto-objects. In the proposed Bayesian approach, states of the target and proto-objects are jointly estimated over time. Gibbs sampling has been used to optimize the estimation during the tracking process. The proposed method robustly handles occlusion, distraction, and illumination change in the experiments. Experimental results also demonstrate that the proposed method outperforms the state-of-the-art methods in challenging tracking tasks.

1. Introduction

Visual tracking, as a fundamental step to explore videos, is important in many computer vision based applications, such as security and surveillance, video compression, and robotic vision systems. During the tracking process, state of the target is estimated over time by associating its representation in the current frame with those in previous frames. Although research on visual tracking has lasted for decades, various noisy factors including background clutter, occlusion, distraction, and illumination variance still cause problems in practice, which makes visual tracking a challenging task [1,2]. Most of the problems will cause unpredictable appearance changes during tracking, especially when the target is a non-rigid object.

In order to improve the robustness of visual tracking, the method of tracking by detection has been proposed by translating visual tracking into classification task [3–5]. Besides, a number of tracking algorithms employ object model updating to deal with appearance changes, such as the incremental learning based target updating in [6] and the on-line sparse principal component analysis in [7]. However, it can be observed that appearance based methods, including those with on-line model updating, may cause drift or even loss of target under relatively complex conditions such as large pose variation and sudden illumination change.

An alternative method is saliency based tracking, which is inspired by biological vision systems [8–10]. Saliency represents the extent of visual attention paid to a place in the scene according to its standing out to surrounding [11]. For saliency detection, a saliency map is built to represent the saliency values of all the regions in the given image [12]. Usually the target is more likely to appear in the places with higher saliency values. Different kinds of methods have been proposed to compute saliency values of an image. Most approaches of saliency detection are based on low-level features [13–17], which focus on detecting salient regions in an image by bottom-up mechanism. In these approaches, different regions are ranked and selected according to the contrast of low-level features between each individual region and its surrounding. Bottom-up saliency detection can be easily applied to various scenarios. However, without task related knowledge, salient regions detected by such methods might be insufficient to comprehensively describe the target or separate the target from the background, as sometimes background areas may also have high contrast with low-level features. Also, some salient regions such as corners and edges could appear and then disappear in a short time when illumination condition or target appearance changes drastically. Therefore, during the tracking process the detected salient regions may lack consistency over time. Other approaches of saliency detection make use of object-level representation through the top-down mechanism [18–20]. That is, with explicit conceptual

* Corresponding author at: National ICT Australia, Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia. Tel.: +61 2 9376 2200.
E-mail addresses: zhidong.li@nicta.com.au (Z. Li), weihong.wang@nicta.com.au (W. Wang), yang.wang@nicta.com.au (Y. Wang), fang.chen@nicta.com.au (F. Chen), yi.wang@nicta.com.au (Y. Wang).

[1] National ICT Australia (NICTA) is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program.

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.020

knowledge about the target, the saliency value over a region measures its chance of belonging to an object, so that the target can be separated from the background based on the object-level saliency. However, such approach is inevitably limited to detecting an already learned target, for example, saliency based human detection in [19]. Compared to bottom-up methods, detecting targets by object-level representation during tracking is usually inconvenient due to the difficulties of obtaining the target conceptual knowledge.

It has been observed that, even when the conceptual knowledge about actual objects is not explicitly presented, humans can still percept a region that is plausible to be an object. Representation level for such kind of plausible objects is between the low-level feature based representation and the object-level semantic representation [21]. In psychology, it can be explained by the coherency theory by Rensink [22] that an amount of attentional regions (regions of the plausible objects) can be selected and bundled together into actual objects. These regions, referred to as proto-objects, are volatile units of visual information that can be accessed by selective attention and subsequently validated as actual objects [22]. Compared to low-level feature based salient regions, proto-objects are determined by simultaneously considering the top-down mechanism and the bottom-up mechanism [21,23,24], so that their correlations to the target are much more stable during the tracking process. In addition, detection of the proto-objects does not require conceptual knowledge about the target.

In this paper, we introduce a framework of visual tracking based on proto-objects. Given an image sequence, proto-objects around the target region are first detected by combining bottom-up saliency and top-down topic model. The target is then tracked over time based on spatial and saliency information of the proto-objects. Sampling based optimization algorithm is utilized to infer the states of both the target and the proto-objects during the tracking process. Experimental results demonstrate that the proposed approach outperforms the state of the art, and it robustly handles occlusion, distraction, as well as illumination variation. To authors' knowledge, the approach of visual tracking by proto-objects is first introduced in this paper.

The paper is organized as follows. Section 2 reviews the related work. Section 3 presents our method for proto-object detection. Section 4 introduces the framework of tracking by proto-objects and the optimization algorithm. Experimental results are discussed in Section 5. The conclusion is drawn in Section 6.

2. Related work

Several saliency based approaches have been proposed for visual tracking. Mahadevan and Vasconcelos proposed salient feature based tracking in [8]. The features for target tracking are selected based on its saliency values, which are estimated according to the power of discriminating the target from its surrounding areas [25]. The technique has also been extended to detect salient regions in spatiotemporal space [26]. In addition, the framework of tracking by salient regions has been proposed in [9,10]. In their work salient regions are first detected and then tracked in video frames based on low-level features using the bottom-up mechanism. Target location is determined through the weighted combination of all available salient regions. Tracking by salient regions often relies on the co-occurrence of a large number of salient regions [10], which is computationally expensive; besides, the target must be large enough [9]. Tracking by proto-objects generally requires less number of proto-objects and it works for small target as well, since the whole target can be detected as salient proto-object.

Yang et al. has proposed an approach of context-aware visual tracking that uses a set of auxiliary objects to support the target tracking [27]. The auxiliary objects are tracked collaboratively with the target during the tracking process. Auxiliary object refers to image region that exhibits persistent co-occurrence and consistent motion with the target. Compared with auxiliary object, proto-object is a biologically inspired concept, and saliency information is used to detect proto-objects in this work. Moreover, in practice most auxiliary objects are detected outside the target region, while proto-objects are usually detected within the target region. Hence the two methods overall utilize different sources of context information, and they could be complementary to each other during the tracking process.

The problem of proto-object detection has also attracted researchers from both psychology and computer vision societies. Existing work on proto-object detection includes the biologically plausible visual attention models [21,23]. For computational implementation, to know how likely a region is a proto-object, the saliency value of the region is determined by the combination of both bottom-up and top-down mechanisms [24]. As an example, proto-objects are detected using the feed-forward and feedback links in [23]. In addition, some work in the field of psychology carried out cognitive experiments to discover various phenomena about proto-objects, such as [22].

3. Salient proto-object detection

Given a video sequence, our method first detects proto-objects in the initial frame based on saliency detection and topic model. For visual tracking, the initialized target region (usually represented by a rectangle bounding box) contains certain knowledge about the target although it is far from comprehensively representing the target. Such knowledge is represented by a top-down topic model in this work. Considering various noisy factors such as appearance or illumination change during the tracking process, the impact of top-down knowledge sometimes becomes weak and implicit. Hence for the robustness of proto-object detection, it is helpful to combine the top-down topic model with the bottom-up saliency.

First, a saliency map is built for the target region and surrounding areas based on low-level features. The low-level features we use consist of color, intensity, and orientation, which are the same as those features used in [12]. For each low-level feature f, a feature map in scale c can be obtained for the image. Then a saliency map is calculated based on the spectral phase information of the feature map [14]. The scale parameter c is determined by the input image size [13], for which we use 32×32, 64×64, and 128×128 in this work. The combination of multi-scale saliency maps helps discover more proto-objects and produces better object boundaries [28]. Hence saliency maps with different features and scales are summed up to form a bottom-up saliency map. Fig. 1b shows an example of saliency map obtained using our method.

Meanwhile, the image is segmented into a set of partitions, and topic model is used to group the partitions with high saliency values to form proto-objects. To ensure that each partition only belongs to one proto-object, we employ an over-segmentation algorithm proposed in [29] (see Fig. 1c for an example). Here we use visual words [30] to represent a partition and PLSA model [31] to group the partitions with consistent representation. The visual words are composed by clustering low-level feature vectors using K-means. The PLSA model assumes that there exists a set of topics for both salient and non-salient regions. An image partition can be assigned to either a salient region topic or a non-salient region topic, while each topic is characterized by certain