Neurocomputing 204 (2016) 87–96

Contents lists available at ScienceDirect

Neurocomputing

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

Deformable object tracking with spatiotemporal segmentation in big vision surveillance

Peng Zhang [a,*], Tao Zhuo [b], Lei Xie [a], Yanning Zhang [a]

[a] Northwestern Polytechnical University, Xi'an, China
[b] National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form
29 June 2015
Accepted 18 July 2015
Available online 9 April 2016

Keywords:
Deformable
Spatiotemporal
Online tracking
Big vision
Surveillance

ABSTRACT

Rapid development of worldwide networks have changed the traditional challenges in vision surveillance to a big data level. Accordingly, the video processing technologies also need to focus more on the new coming big vision problems such as efficient content understanding. As a fundamental and indispensable pre-step for high-level video analysis, e.g. behavior recognition for social security, accurate and robust object tracking can play an essential role because of its capability in extracting the salient information from the captured video dataset. Due to the complexity of the realistic application environments, accurate and robust tracking is not easy because the object appearance may continually change during its moving, especially for the deformable objects, it is difficult for the designed appearance model being adaptive to the heavy shape variations as rotation or distortion. In this paper, a novel object tracking based on spatial segmentation is proposed to handle the problem of drastic appearance changes of the deformable object. By using the motion information between the consecutive frames, the irregular areas of the deformable object can be segmented more accurately by energy function optimization with boundary convergence. In succession, the segmentation areas are modeled by a structural SVM as learning samples to achieve more effective online tracking. Based on the evaluation of the proposed tracking on the standard benchmark database containing the challenges of heavy intrinsic variations and occlusions, the experiment results have demonstrated a significant improvement in accuracy and robustness when compared with other state-of-art tracking approaches.

1. Introduction

In the past decades, different types of video surveillance systems have demonstrated their effectiveness for public security all around the world. With the rapid development of the HDTV and mobile networks nowadays, the video data volume and resolution having been enhanced with an incredible speed, which makes most of the current surveillance system [1,2] need to face the similar challenges introduced by big data analytics as in the other areas as data storage or information retrieval for the obtained big videos [3]. In recent years, the content based learning-retrieval [4] mechanism has been verified a promising solution for the ‘big’ video data analysis, for which the learning sample cannot be extracted by offline as we usually do on the normal video dataset. Thus, employing some automatic object tracking to obtain the salient features for learning sample generation is a common way in many learning strategies. Specifically, more accurate tracking for learning sample generation, more satisfactory the retrieval result. Therefore, the research of visual tracking is still an essential topic in handling the challenge problems in the coming big data age.

Tracking a rigid shape object in a simple surveillance environment, such as a car running on a highway, has been resolved in different satisfactory ways [5–8]. However, tracking the deformable object in realistic scenarios is still hard because the target appearance may change constantly during moving [9], especially for an irregular-shaped object, challenges mainly come from the intrinsic variations such as distortion, rotation and scaling [10]. In order to effectively adapt to such target appearance changes, the most popular way of conducting online tracking is to update the appearance model and make it suitable for distinguishing the object from background on-the-fly. In general cases, due to the deficiency knowledge to the learner before learning starts, the noisy constraints in training data leads to performance degeneration when updating classifiers. Hence, in order to introduce more rational paradigms between training data for classifier updates, Kalal et al. proposed a P–N learning scheme and applied it to the problem of online tracking-by-learning [11]. P–N learning

[*] Corresponding author.
E-mail addresses: zh0036ng@nwpu.edu.cn (P. Zhang), zhuotao@nus.edu.sg (T. Zhuo), lxie@nwpu.edu.cn (L. Xie), ynzhang@nwpu.edu.cn (Y. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.07.149
0925-2312/© 2016 Elsevier B.V. All rights reserved.

establishes a novel structure of the training samples by exploiting the positive and negative constraints, which restricts the data labeling operation [3]. This framework also helps to guide the design of more sophisticated structural constraints that can fulfill requirements of the learning stability. However, its limitation resides in the usage of inaccurate positive data sample which sampled in the background areas inside the target bounding box. In addition, tracking failure in many cases is still hard to be avoided because of the inaccurate motion estimation between unreliable consecutive frames.

Similar as Kalal's tracker, many other proposed online tracking strategies also utilized regular geometry shapes such as bounding-box or ellipse to represent the appearance of the target. These regular-shape based tracking methods can track the targets of fixed shape such as human head or cars robustly, but the tracking failure often happen in handling the irregular-shaped targets with good accuracy, especially when targets have heavy partial occlusions or intrinsic variations. To overcome this limitation, Kwon et al. proposed an approach by using a pre-defined bounding-box collection to represent different parts of the target [12] for articulated object tracking. With the help of an adaptive Basin Hopping Monte-Carlo Sampling (BHMC) method, Kwon's approach can automatically update the target dynamic appearance changes and geometry relations over time. Likely, Yao et al. also utilized a global object box and a set of part boxes as an appearance model to approximate the irregular-shaped object [13]. With an online two-stage training mechanism to learn the parameter of part based model, Yao's strategy is able to overcome the complexity problem due to model overfitting. Different from the approaches [12] and [13] using regular part-based representation for single target, Zhang et al. introduced a model-free tracker that simultaneously tracks multiple objects by combining multiple single-object trackers via constraints on the spatial structure of the objects. The performance of this structure-preserving tracking approach show an obviously improvement in multi-object tracking by using an online structured SVM algorithm, which is similar as [13].

In many realistic situations, unfortunately, such pre-defined parts representation influences the extendability and generic application for those methods [14], especially when the target is composed by several objects, e.g. motor rider shown as Fig. 5, which is difficult to be effectively represented by discrete rectangles. Thus, more accurate representation such as using continuous contour to smoothly estimate target's contour/shape would be a possible way for target presentation. Sun et al. proposed a supervised level set model for tracking [15] in order to obtain more precise convergence to the target during tracking. With the specific knowledge of target region and edge cues, the contour curve can converge to the candidate area with maximum likelihood in a Bayesian manner. Recently, Hough-transform based approaches have received attraction in overcoming the limitations of using fixed-shape set for irregular representation [16]. Barinova et al. proposed a probabilistic framework for multiple object instances detection in Hough transforms domain [17,18]. And the main point of this research also inspired the following work proposed by Godec et al. [10]. In Godec's tracker, the GMM based segmentation [19] is incorporated with Hough forest learning framework for irregular-shaped target tracking [16]. With the back-projection to support an online tracking process, Godec's tracker beyond many state-of-art work based on fixed-shape appearance representation. However, heavily relying on the discrimination of color Gaussian kernels makes the segmentation result unexpectedly, especially when intrinsic changes happen in the uniform color background containing obvious edges.

Aiming at the problem of learning data generation, in this paper, we propose a novel motion-appearance model to achieve accurate spatiotemporal segmentation for deformable object online tracking. Compared with the spatial guided segmentation [10] only depending on the texture/pixel information within the individual frame, the proposed model is able to segment the target areas more precisely with the help of motion information between consecutive frames, especially when the texture of background and target are similar. The segmented areas can provide more precise samples for online model updating. To effectively describe the appearance of deformable object, we utilize the structural SVM [20] to construct an online *tracking* framework, which is more accurate than only employing fixed rectangle for target appearance modeling in spatial domain. The proposed tracking shows more robust in many challenge scenes including rotation, intrinsic compression/stretching and aspect ratio changes, for vision based surveillance application.

The following organization of the paper is as: Section 2 introduces the proposed spatiotemporal segmentation with motion-appearance model and Section 3 briefly introduced the online learning tracking framework. Section 4 shows experiment results and discussion and Section 5 gives the conclusion of this paper.

2. Learning samples generation by spatiotemporal segmentation

Learning samples generation is an important issue for most of online tracking-by-learning strategies [11]. For the fixed-shape based approaches, the main challenge is: the correctness of generated samples for online learning is hard to be guaranteed due to the noise in coarse data by the annotated bounding boxes. The purpose of the proposed segmentation model is to separate the foreground from the background with fine-smooth contour. The accurately separated foreground areas can provide higher quality samples for online learning. The segmented region are consisted of two parts: positive samples which are within the segmented foreground areas; negative ones which are in the background. During initialization, the occupied areas by target are positive while the other positions inside the maximum-object-sized bounding-box are regarded as background.

2.1. Notation definitions

In the beginning of this section, we firstly list out all the notations and the acronyms we have used in the following content for easy reference of the readers.

$\{f_1, \ldots, f_t\}$ – a video frame sequence	$\mathbf{p}(x, y, t)$ – image lattice for f_t
$\{f(\mathbf{p}_1), f(\mathbf{p}_2), \ldots, f(\mathbf{p}_N)\}$ – N-size array	$\{\alpha_1, \ldots, \alpha_N\}$ – an array of 'opacity' values
$\underline{\theta}$ – foreground and background grey-level distributions	H – normalized histogram model
E_φ – energy function of opacity distribution	E_ϕ – energy function of boundary coherence
E_ψ – energy function of piecewise smoothness in flow field	$\oplus$ – interleaved Optimization
C – boundary neighborhood region	γ – smoothness regularization parameter
β – smoothness regularization parameter	$s(\cdot)$ – Euclidean distance
$\mathbf{w}$ – displacement flow field	U, V, dU, dV – vectorization of u,v,du,dv