Computer Vision and Image Understanding 120 (2014) 70–80

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Efficient and robust multi-template tracking using multi-start interactive hybrid search ☆

Hadi Firouzi *, Homayoun Najjaran

Okanagan School of Engineering, The University of British Columbia, Kelowna, BC, Canada

ARTICLE INFO

Article history:
Received 25 August 2012
Accepted 20 December 2013
Available online 30 December 2013

Keywords:
Adaptive template
Gaussian-based appearance model
Multi-model representation
Parallel optimization
Non-rigid object

ABSTRACT

This paper presents an efficient, accurate, and robust template-based visual tracker. In this method, the target is represented by two heterogeneous and adaptive Gaussian-based templates which can model both short- and long-term changes in the target appearance. The proposed localization algorithm features an interactive multi-start optimization process that takes into account generic transformations using a combination of sampling- and gradient-based techniques in a unified probabilistic framework. Both the short- and long-term templates are used to find the best location of the target, simultaneously. This approach further increased both the efficiency and accuracy of the proposed tracker. The contributions of the proposed tracking method include: (1) *Flexible multi-model target representation* which in general can accurately and robustly handle challenging situations such as significant appearance and shape changes, (2) *Robust template updating algorithm* where a combination of tracking time step, a forgetting factor, and an uncertainty margin are used to update the mean and variance of the Gaussian functions, and (3) *Efficient and interactive multi-start optimization* which can improve the accuracy, robustness, and efficiency of the target localization by parallel searching in different time-varying templates. Several challenging and publicly available videos have been used to both demonstrate and quantify the superiority of the proposed tracking method in comparison with other state-of-the-art trackers.

1. Introduction

Efficiency and reliability of many computer vision applications such as automated visual surveillance systems [1], motion analysis and activity recognition [2,3], vehicle navigation and tracking [4], and intelligent preventive safety systems [5] highly depend on their visual object tracking algorithm. In spite of extensive research, object tracking is still challenging generally due to the inevitable object appearance changes, scale variations, different lighting conditions, noise and outliers, complex motion, and cluttered environments. Specifically, the main difficulty in tracking non-rigid objects – which has been emphasized in this work – is related to the high dimensional complexity and uncertainty in real-world applications [6].

In its simplest form, visual tracking can be viewed as the problem of locating and corresponding one or more image regions to a specific object in an image sequence. Visual trackers are roughly classified as either direct (image-based) methods or indirect (feature-based) methods. In the latter approach, different feature descriptors including silhouette [7], contour [8], texture [9], local invariant features [10], Haar-like features [11], and histograms [12,13] are used to model the object appearance. Feature descriptors can – to some degree – handle illumination changes, scale and appearance variations, and outliers. However, their suitability and robustness may significantly change from one application to another depending on the appropriateness of the feature descriptors used. Moreover, those descriptors such as SIFT [10], which can suitably handle the object non-rigidity and scale variations in many cases, are hampered in real-world applications due to the high computational cost of feature extraction and matching algorithms. On the other hand, in direct methods, the object appearance is modeled by one or more sub-images, known as *templates*, which consist of the image pixel values. Object regions are composed of both appearance and spatial information so that direct methods can track objects more accurately and robustly.

Template matching is a well-studied computer vision problem which has been introduced by [14] for the task of visual tracking. In this method, the target region is considered as the template $T(X)$ where $X = (x, y)$ is the pixel coordinate, and the goal is to find the best corresponding match in the next image I^n based on the target dynamical model $W(X;\Theta)$ where $\Theta = \{\theta_1, \ldots, \theta_k\}$ are the template transformation parameters. By minimizing the equation:

$$\Theta^n = \arg\min_{\Theta} \sum_{X \in T} [I^n(W(X;\Theta)) - T(X)]^2 \qquad (1)$$

☆ This paper has been recommended for acceptance by Nikos Paragios.

* Corresponding author.

E-mail addresses: hadi.firouzi@ubc.ca (H. Firouzi), h.najjaran@ubc.ca (H. Najjaran).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.12.009

which is the sum of squared differences (SSD) between the template $T(X)$ and the candidate sub-image $I^n(W(X;\Theta))$ we can find the best match transformation parameters Θ^n. A gradient-based algorithm to optimize Eq. (1) was introduced in [14]. However, conventional template trackers are not generally robust to significant appearance and scale variations, occlusion, and illumination changes.

To address the shortcomings of template matching, in this paper a robust template tracking method based on the Sum of Gaussian Errors (SGE) between the object template and the candidate sub-image is proposed. The object template is modeled by several Gaussian functions which are adaptively updated to handle both the object appearance changes and the "drift" problem.[1] At every tracking step, a certain number of probabilistic optimization processes with different starting points are performed in parallel to estimate the object location. The proposed method is capable of tracking non-rigid objects with variable appearance, shape, and scale in cluttered environments. It is assumed that the object location is specified either manually or automatically (by any existing object detection method) at the beginning and the main goal is to track the object without any prior knowledge about the object appearance and motion dynamic.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. In Section 3, the proposed multi-model target representation is defined. The formulation and algorithm of the proposed multi-start Gaussian-based template tracking method are explained in details in Section 4. In Section 5, the proposed tracker is applied on five challenging image sequences and subsequently the results are compared with four state-of-the-art methods as well as the ground truth data. Concluding discussions and potential extensions for future work will be provided in Section 6.

2. Related work

Since early template-based tracking methods [14], different algorithms have been proposed to improve the accuracy and efficiency of the tracking; Bergen et al. [16] used a more general motion model e.g., affine transformation, Black and Jepson [17] improved the robustness of the template matching against appearance changes by employing a linear appearance variation, Hager and Belhumeur [18] increased the tracking efficiency by a real-time implementation, and Cootes et al. [19] modeled the object appearance by Active Appearance Models (AAMs) to handle non-rigid objects. Matthews et al. [15] proposed a method based on an adaptive appearance template which does not suffer from the "drift" problem. Instead of using previous update strategies which involve either no update ($T_{n+1} = T_1$ for all $n \geqslant 1$) or a naive template update ($T_{n+1} = I_n(W(X;\Theta_n))$ for all $n \geqslant 1$), they first estimate new transformation parameters Θ_{n+1} based on the naive template update, and then the estimated parameters are used as a starting point to align template T_{n+1} with T_1. This method is relatively stable to the local minimum by reinitializing the gradient-based search. However, the method proposed in [15] cannot handle the occlusion and outliers, and it also fails when tracks non-rigid objects, especially when the object shape is changing over time. Schreiber [20] presented a robust template matching algorithm to handle partial occlusion and outliers. Unlike other robust template trackers such as [18,21], in this method the robust weights are adaptively updated only after finding the transformation parameters for a new image to improve the computational efficiency. However, according to the experimental results, this method is mainly robust to track rigid objects under different lighting conditions and partial occlusion. Silveira and Malis [22] used several image transformation models to improve the template-based tracking performance against illumination changes. They proposed a new illumination model which can be used to track a deformable target with illumination change. In this method, a general image formation model which covers both geometric and photometric deformations is defined to track a rigid or deformable object. Although the proposed method is robust to illumination change and general object deformation, it cannot handle large and unpredicted pose and appearance changes due to the gradient-based optimization and a large number of parameters estimated. Also this method is not stable when the target image is not sufficiently textured and unable to track non-rigid objects with variant shape and structure.

Recently Firouzi and Najjaran [23] presented a component-based template tracking method using a multi-start EM-like localization algorithm. Building on their ideas, this paper describes an overhaul of the approach in [23] to improve its performance and robustness for real-time applications. The contributions and differences of the present work in comparison with the work published in [23] include:

- *Target representation model*: In the method presented in [23], the image region of the target object is initially partitioned into several sub-regions, and subsequently each sub-region is represented by two Gaussian-based templates namely immediate and delayed templates. Similarly, the proposed representation model consists of two time-varying templates which can model both short-term and long-term changes in the target appearance. However, other appearance models such as Local Binary Pattern (i.e., a texture descriptor insensible to illumination changes) can be efficiently integrated into the proposed multi-model target representation which is not possible in [23]. Considering the structural differences, the current target model can be customized suitably to obtain a more satisfactory result in comparison with the model presented in [23].
- *Representation model learning*: The mean and variance of the templates used in [23] are updated separately based on an updating ratio and the tracking time step, respectively. On the other hand, both the mean and variance of the templates in the current work are adaptively updated based on a forgetting factor, uncertainty margin, and the tracking time step. Therefore, in comparison with the method presented in [23] the proposed template update strategy is not only more adaptive to new appearance changes because of the use of a forgetting factor and the tracking time step, but also more robust against noise and occlusion due to the uncertainty margin used in the learning algorithm.
- *Target localization algorithm*: The method presented in [23] uses a predefined multi-start Gradient-based search to estimate the preliminary location of the target sub-regions based on a translational transformation and immediate template. Consequently, the delayed template is employed to correct the preliminary estimation. At the end of this two-step optimization, the target is tracked by fusion of the new sub-region locations. In contrast, the target localization in the current work features an interactive multi-start hybrid search that takes into account generic transformations using a combination of sampling- and gradient-based algorithms in a unified probabilistic framework. Unlike the two-step optimization used in [23], in the current method all appearance models (i.e., the short- and long-term templates) are used to find the best location of the target, simultaneously. This approach further increased both the efficiency and accuracy of the proposed tracker.

[1] The problem of gradually updating the object appearance model with irrelevant information such as background pixel values [15].