Image and Vision Computing 33 (2015) 44–56

Contents lists available at ScienceDirect

Image and Vision Computing

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

Robust tracking with interest points: A sparse representation approach ☆

R. Venkatesh Babu *, Priti Parate, Äniruddha Acharya K.

Video Analytics Laboratory, Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore, India

ARTICLE INFO

Article history:
Received 6 December 2013
Received in revised form 14 July 2014
Accepted 21 October 2014
Available online 7 November 2014

Keywords:
Visual tracking
l_1 minimization
Interest points
Harris corner
Sparse representation

ABSTRACT

Visual tracking is an important task in various computer vision applications including visual surveillance, human computer interaction, event detection, video indexing and retrieval. Recent state of the art sparse representation (SR) based trackers show better robustness than many of the other existing trackers. One of the issues with these SR trackers is low execution speed. The particle filter framework is one of the major aspects responsible for slow execution, and is common to most of the existing SR trackers. In this paper,[1] we propose a robust interest point based tracker in l_1 minimization framework that runs at real-time with performance comparable to the state of the art trackers. In the proposed tracker, the target dictionary is obtained from the patches around target interest points. Next, the interest points from the candidate window of the current frame are obtained. The correspondence between target and candidate points is obtained via solving the proposed l_1 minimization problem.
In order to prune the noisy matches, a robust matching criterion is proposed, where only the reliable candidate points that mutually match with target and candidate dictionary elements are considered for tracking. The object is localized by measuring the displacement of these interest points. The reliable candidate patches are used for updating the target dictionary. The performance and accuracy of the proposed tracker is benchmarked with several complex video sequences. The tracker is found to be considerably fast as compared to the reported state of the art trackers. The proposed tracker is further evaluated for various local patch sizes, number of interest points and regularization parameters. The performance of the tracker for various challenges including illumination change, occlusion, and background clutter has been quantified with a benchmark dataset containing 50 videos.

1. Introduction

Visual tracking has been one of the key research areas in computer vision community for the past few decades. Tracking is a crucial module for video analysis, surveillance and monitoring, human behavior analysis, human computer interaction and video indexing/retrieval etc. Major challenges for tracking algorithms that arise in real life scenarios are due to both intrinsic and extrinsic factors. Intrinsic factors include pose, appearance and scale changes, and common extrinsic factors are illumination variation, occlusion and clutter.

There have been several proposals for object tracking algorithms in the literature. Based on object modeling, a majority of the trackers can be brought under the following two themes: i) Global object model and ii) Local object model. In the global approach, the object is typically modeled using all the pixels corresponding to the object region or some global property of the object. The simple template based SSD (sum of the squared distance) tracker, color histogram based meanshift tracker [1] and probabilistic tracker [2] are examples of global trackers. Traditional Lucas–Kanade tracker [3], fragment based approaches [4,5] and many bag-of-words model based trackers are some examples of local object trackers.

In global modeling, the features representing the global properties of the object are utilized for modeling. They could be simple template based models or histogram based models or shape based models. The template models carry appearance information of the object from a single view. These models are good for tracking objects whose appearances do not change much over time and not suitable for tracking objects undergoing significant appearance changes, which need frequent model updates. Since image intensity based template models are sensitive to illumination changes, image gradients have been used as a feature [6]. Template matching approach is computationally very expensive due to the brute force search. Efficient template matching methods have been proposed in the literature [7,8]. On the other hand, Comaniciu et al. [1] use the kernel weighted color histogram for modeling the object. Though the spatial information is lost in this model, it is suitable for applying iterative meanshift procedure. This meanshift tracker maximizes the similarity between the target and candidate models by iteratively seeking the mode of the underlying similarity space. Since the meanshift tracker performs gradient ascent over similarity space, it quickly converges to the mode in a couple of iterations and delivers real-time tracking

☆ This paper has been recommended for acceptance by Ming-Hsuan Yang.

* Corresponding author.
E-mail address: venky@serc.iisc.in (R. Venkatesh Babu).
URL: http://www.serc.iisc.ernet.in/~venky/ (R. Venkatesh Babu).

[1] An earlier brief version of the paper has appeared in ICIP'13 (R. Venkatesh Babu and P. Priti, "Interest points based object tracking via sparse representation", in proceedings of International Conference on Image Processing (ICIP), Melbourne, Australia, 2013).

http://dx.doi.org/10.1016/j.imavis.2014.10.006
0262-8856/

performance. Unlike the template based object model, the histogram based object model provides better robustness to appearance change, since it captures the color configuration of the object rather than the spatial structure. These global object modeling approaches are sensitive to partial occlusion, illumination and scale changes since the model depends on the attributes of entire object region.

On the other hand, local object model uses the information from object parts for tracking. Most of these trackers use bag-of-words model based approaches [9–12,5]. The proposed interest point based tracker falls under the local object model based tracking. Shi and Tomasi showed [13] that the corner-like points are more suitable for reliable tracking due to its stability and robustness to various distortions like rotation, scaling, and illumination. Though the interest-point based trackers show more robustness to various factors like rotation, scale and partial occlusion, the major issues surface from description and matching of interest points between the successive frames. For example, Kloihofer and Kampel [14] use SURF [15] descriptors of interest points as feature descriptors. The object is tracked by matching the object points of the previous frame with candidate points in the current frame. The displacement vectors of these points are utilized for localizing the object in the current frame [16]. A detailed survey on various object tracking methods can be found in [17–19]. Matching the features of interest points between two frames is a crucial step in estimating the correct motion of the object and typically Euclidean distance is used for matching. The recently proposed vision algorithms in sparse representation framework clearly illustrate its superior discriminative ability even with very low dimensional data [20,21]. This motivated us to examine the matching ability of sparse representation approach for interest points.

In this paper, we have proposed a robust interest point based tracker in sparse representation framework. The interest points of the object are obtained from the initial frame by Harris corner detector [22] and a dictionary is constructed from the small image patch surrounding these corner points. The candidate corner points obtained from the search window of the current frame are matched with object points (dictionary) by sparsely representing the candidate corner patches in terms of dictionary patches. The correspondence between the target and candidate interest points is established via the maximum value of the sparse coefficients. A 'robust matching' criterion has been proposed for pruning the noisy matches by checking the mutual match between candidate and target patches. The displacement of these matched candidate points indicates the location of the object. Since the dictionary elements are obtained from a very small patch surrounding these corner points, the proposed approach is robust and computationally very efficient compared to the particle filter based l_1 trackers [23–25].

The rest of the paper is organized as follows: Section 2 briefly reviews related works. Section 3 explains the proposed tracker in sparse representation framework. Section 4 discusses the results and concluding remarks are given in Section 5.

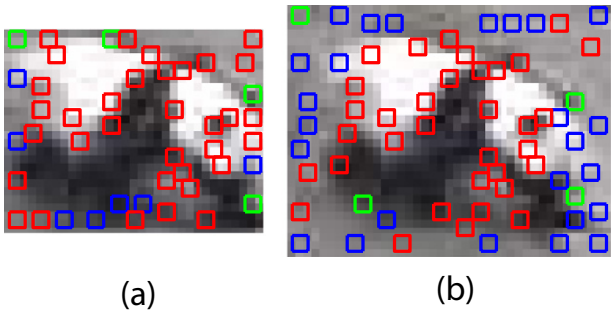

Fig. 2. Robust matching of interest points. (a) Target window with interest points (b) Candidate window with interest points. The red patches are the mutually matched points, green patches matched only one way (either target to candidate or candidate to target) and the blue patches are unmatched ones.

2. Sparse representation based tracking

The concept of sparse representation recently attracted the computer vision community due to its discriminative nature [20]. Sparse representation has been applied to various computer vision tasks including face recognition [20], image video restoration [26], motion segmentation [27], image denoising [28], image compression [29], action recognition [30], super resolution [31], tracking [21] and background modeling [32].

Wright et al. [20] exploited the discriminative nature of sparse representation for face recognition. In place of generic dictionaries, they have used the overcomplete dictionary with the training samples as its base elements. Given sufficient number of training samples for each class, the test sample can be sparsely represented using the training elements of the same class via l_1 minimization. The same concept can be effectively used for object tracking, since the correct candidate can be sparsely represented using target dictionary.

The l_1 minimization tracker proposed by Mei and Ling [33] uses the low resolution target image along with trivial templates as dictionary elements. The candidate patches can be represented as a sparse linear combination of the dictionary elements. To localize the object in the future frames, the authors use the particle filter framework. Here, each particle is an image patch obtained from the spatial neighborhood of previous object center. The particle that minimizes the projection error indicates the location of object in the current frame. Typically, hundreds of particles are used for localizing the object. The performance of the tracker relies on the number of particles used. This tracker is computationally expensive and not suitable for real-time tracking. Faster version of the above work was proposed by Bao et al. [24], here the l_2 norm regularization over trivial templates is added to the l_1 minimization problem. However, our experiments show that the speed of the

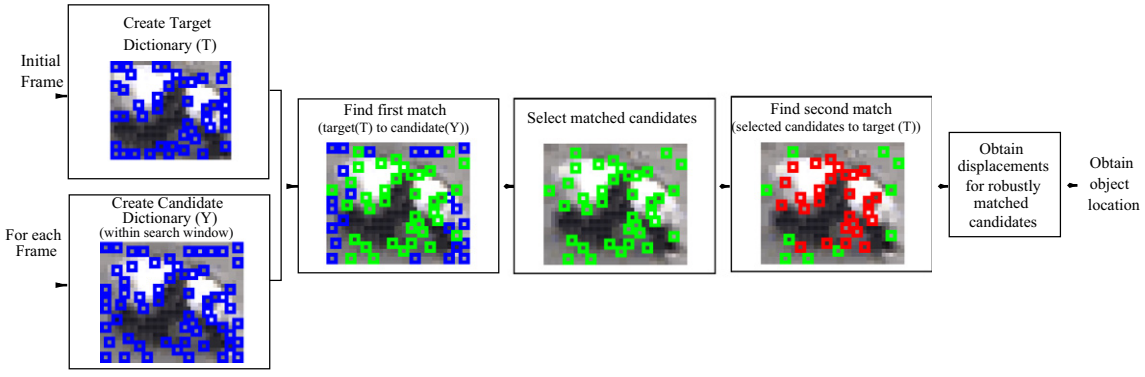

Fig. 1. Proposed tracker overview. The 'blue' patches indicate the initial target and candidate interest points. The 'green' patches indicate the '1-way' matched points and the 'red' patches indicate the robustly matched ('2-way' matching) interest points.