J. Vis. Commun. Image R. 56 (2018) 1–14

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Robust visual tracking via multi-feature response maps fusion using a collaborative local-global layer visual model ☆

Haoyang Zhang [a,b], Guixi Liu [a,b,*], Zhaohui Hao [a]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China
[b] Shaanxi Key Laboratory of Integrated and Intelligent Navigation, Xi'an, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 26 September 2017
Revised 7 July 2018
Accepted 21 August 2018
Available online 23 August 2018

Keywords:
Collaborative visual model
Block color tracking
Correlation filter tracking
Response maps fusion
Online re-detection

ABSTRACT

This paper addresses the issue of robust visual tracking, in which an effective tracker based on multi-feature fusion under a collaborative local-global layer visual model is proposed. In the local layer, we implement a novel block tracker using structural local color histograms feature based on the foreground-background discrimination analysis approach. In the global layer we implement a complementary correlation filters-based tracker using HOG feature. Finally, the local and global trackers are linearly merged in the response maps level. We choose the different merging factors according to the reliability of each combined tracker, and when both of the combined trackers are unreliable, an online trained SVM detector is activated to re-detect the target. Experiments conducted on challenging sequences show that our final merged tracker achieves favorable tracking performance and outperforms several state-of-the-art trackers. Besides, performance of the implemented block tracker is evaluated by comparing with some relevant color histograms-based trackers.

1. Introduction

Visual tracking is an important research topic for many computer vision tasks including robotics, surveillance, and human-computer interaction, to name a few [1]. By specifying the target in the initial frame, the kernel problem of visual tracking is to continuously estimate the best configuration of the target in the coming frames. Over the past decade, numerous tracking methods have been proposed and significant progress has been made, yet it is still challenging for an existing tracker to simultaneously deal with the complicated situations such as illumination variations, occlusion and shape deformation [2,3]. To address this issue, a more comprehensive and robust tracker is needed.

Generally speaking, tracking methods can be classified as either generative or discriminative tracking based on the appearance models used [3]. Generative methods [4–8] learn a generic appearance model according to the given templates or subspace models of the target, and tracking is accomplished by performing the searching operation for the best matching score within the target region. Discriminative methods take advantage of the appearance information of the target and background, and treat the tracking as a binary classification process (known as tracking-by-detection [9]) that aims at distinguishing the target from the background [10–12]. Recently, discriminative correlation filters-based (DCFs) approaches have drawn extensive attention in computer vision fields and been successfully applied in visual tracking [13–19]. One of the prominent merits that highlights the DCFs-based visual tracking among others discriminative approaches is that the DCFs are very efficient in training and detection stage as they can be transferred into the Fourier domain and operated in element-wise multiplication, which is of significance for the real-time tracking.

Despite the development of visual tracking in model representation, such as sparse representation and correlation filters, with some well-engineered features like HOG and CN (color names) [16], the previous work [20] still argued that trackers based on standard color representations can achieve competitive performance. Furthermore, in [21], Bertinetto et al. proposed the Staple (sum of template and pixel-wise learners) tracker by complementing the color histograms-based models with DCFs, and showed that a simple linear combination of response maps of these two models gets a marvelous success in visual tracking. Following the seminal work of [15,20,21], in this paper we implement an effective tracking method by proposing a collaborative local-global layer tracking framework. Our local tracker is based on multiple overlapped local target patches where each patch is represented by the color his-

☆ This paper has been recommended for acceptance by Radu Serban Jasinschi.

* Corresponding author at: School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China.PR China

E-mail address: gxliu@xidian.edu.cn (G. Liu).

https://doi.org/10.1016/j.jvcir.2018.08.018
1047-3203/© 2018 Elsevier Inc. All rights reserved.

tograms. By analyzing the foreground-background discrimination of each patch, the weight of each patch can be determined. The global tracker, as a supplement for our local tracker based on color feature, is a DCFs-based tracker where the HOG feature is been applied. The tracking results of local and global tracker are combined in the response maps level, and we implement a conditional fusion strategy by analyzing the reliability of each combined tracker. To achieve a more robust tracking performance, a SVM classifier is trained and updated during the tracking. The SVM detector can be used to determine the reliability of the local color tracker, and also to perform a re-detect process when both of the combined trackers are unreliable. The overall flowchart of our tracking method is shown in Fig. 1, and the main contributions of this paper are summarized as follows:

(1) The application of foreground-background discrimination analysis method in weighting the different parts of the target, by which a structural local color model-based (SLC) tracker is proposed.
(2) Implement a novel conditional fusion strategy to combine the two response maps from the SLC tracker and global DCFs-based tracker, by which the final local-global confidence maps fusion (LGCmF) tracker is formulated.
(3) A thorough performance comparison of the proposed trackers with several state-of-the-art trackers based on the experiments performed on 80 challenging sequences.

2. Related work

2.1. Color histograms-based visual model

Color histograms have been used in many visual tracking approaches [20–26]. One of the most important properties of color histograms is its insensitivity to shape variation, which is of significance for tracking non-rigid objects. An early implementation of color histograms in visual tracking is the Meanshift tracking [22], where the target position is located by minimizing the Bhattacharyya distance of color histograms between the target and the candidate area using the Meanshift iteration. Abdelai et al. [23] combined Bhattacharyya kernel and integral image as a similarity measure to find the image region most similar to the target. In [20,21,24–26], the histogram model of the target is applied to produce the backprojection map of the searching area, on which each value reflects the probability of corresponding location belonging to the target. Especially, Possegger et al. [20] proposed a discriminative color model by analyzing the target distractors during training and detection stage, according to which the novel DAT (distractor-aware tracker) was formulated. Duffner et al. [25] applied the backprojection maps to segment the target from the background region, and the visual tracking is accomplished by combining a local Hough-Voting model.

Given the two regions R_f and R_s, where the first is corresponding to the foreground of the target and second to its surroundings. We denote H_f and H_s as the color histograms of the two regions. For a pixel $I(x)$ at location x in frame I, the normalized likelihood that $I(x)$ corresponds to the foreground and surroundings can be depicted as:

$$P(x|R_f) = \frac{H_f(I(x))}{|R_f|} \quad \text{and} \quad P(x|R_s) = \frac{H_s(I(x))}{|R_s|} \tag{1}$$

Here $|R_f|$ denotes the number of pixels in the region R_f, $H_f(I(x))$ represents the bin value of the pixel $I(x)$ in the histogram H_f. We denote Γ as the target color model. The probability that a pixel at location x belongs to the target Γ can then be induced by Bayes rule as follows [26]:

$$P(\Gamma|x) = \frac{P(x|R_f)P(R_f)}{P(x|R_f)P(R_f) + P(x|R_s)P(R_s)} \tag{2}$$

the prior probability can be approximated as:

$$P(R_f) = \frac{|R_f|}{|R_f| + |R_s|} \quad \text{and} \quad P(R_s) = \frac{|R_s|}{|R_f| + |R_s|} \tag{3}$$

submit Eqs. (1) and (3) to Eq. (2) and we get:

$$P(\Gamma|x) = \frac{H_f(I(x))}{H_f(I(x)) + H_s(I(x))} \tag{4}$$

2.2. Part-based visual models

Part-based appearance models address the local information of the object, and hence have been proved to be particularly useful in dealing with part occlusion and deformation [27–34]. Adam et al. [27] depicted the target template by arbitrary fragments. The possible states of the target at current frame were voted by each patch according to its histogram similarity with the corresponding target patch. Jia et al. [28] proposed the structural local sparse appearance model, where a novel alignment-pooling method was utilized to exploit the feature of the target. Kwak et al. [29] divided the target into regular grid cells and employed a binary classifier to learn the occlusion status of different parts of the target. When the learned occlusion model is accurate, this method can efficiently handle the occlusion as it obtains the specific occlusion state of the target, while training such a precise occlusion model requires enough training data, which is not always feasible in practice. Kwon et al. [30] successfully addressed the problem of appearance deformation by incorporating a flexible star-like model with a Bayesian filtering framework. However, this method focuses on local appearance of the target while ignoring its whole visual information, and hence it is prone to drift under the scenes like motion blur and clutter. Some authors emphasize the different importance of each patch in the part-based models. For example, Lee et al. [31] selected only pertinent patches that occur repeatedly near the center of the target to construct the foreground appearance model. Li et al. [32] voted the target state by the reliable patches that are exploited according to the trackability and motion similarity of each patch. Recently the coupled-layer visual model has been applied in the visual tracking [33,34]. In these trackers, the local and global information of the target appearance are taken into account simultaneously and are combined in a coupled way, hence achieve more robust tracking performance compared to those trackers using only local information.

2.3. Discriminative correlation filters with multi-channel features

The DCF presented in [15] learns the convolution filter from an image patch $\mathtt{x}$ with $M \times N$ pixels extracted from the center of the target. The training samples are generated by all the cyclic shifts of $\mathtt{x}: \mathtt{x}(m,n), (m,n) \in \{0,\ldots,M-1\} \times \{0,\ldots,N-1\}$. Consider representing these samples by some multi-channel features such as HOG, and then a multi-channel filter f can be learned by minimizing the squared error cost function as follows:

$$\varepsilon(f) = \left\| \sum_{l=1}^{d} \mathtt{x}^{(l)} * f^{(l)} - g \right\|^2 + \lambda \sum_{l=1}^{d} \|f^{(l)}\|^2 \tag{5}$$

Here, $*$ denotes circular convolution. g is a scalar valued function of size $M \times N$ with its values $g(w,h)$ represent the desired convolution response output corresponding to the training samples $\mathtt{x}(w,h)$. $l \in \{1,\ldots,d\}$ is the feature dimension number of $\mathtt{x}$, and $\lambda \geqslant 0$ is a regulation parameter. As the desired response g is constructed as a Gaussian function, the coefficients of the learned filter f are mod-