Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Robust Visual Tracking via Local-Global Correlation Filter

Heng Fan,[1] **Jinhai Xiang**[2,*]
[1]Computer & Information Sciences Department, Temple University, Philadelphia 19122, USA
[2]College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China
hengfan@temple.edu, jimmy_xiang@mail.hzau.edu.cn

Abstract

Correlation filter has drawn increasing interest in visual tracking due to its high efficiency, however, it is sensitive to partial occlusion, which may result in tracking failure. To address this problem, we propose a novel local-global correlation filter (LGCF) for object tracking. Our LGCF model utilizes both local-based and global-based strategies, and effectively combines these two strategies by exploiting the relationship of circular shifts among local object parts and global target for their motion models to preserve the structure of object. In specific, our proposed model has two advantages: (1) Owing to the benefits of local-based mechanism, our method is robust to partial occlusion by leveraging visible parts. (2) Taking into account the relationship of motion models among local parts and global target, our LGCF model is able to capture the inner structure of object, which further improves its robustness to occlusion. In addition, to alleviate the issue of drift away from object, we incorporate temporal consistencies of both local parts and global target in our LGCF model. Besides, we adopt an adaptive method to accurately estimate the scale of object. Extensive experiments on OTB15 with 100 videos demonstrate that our tracking algorithm performs favorably against state-of-the-art methods.

1. Introduction

Visual tracking plays a crucial role in computer vision and has a variety of applications such as robotics, surveillance, human-computer interaction and so forth (Wu, Lim, and Yang 2015). Despite great progress in recent years, object tracking remains a challenging problem due to appearance variations caused by partial occlusion, illumination changes, deformations and so on. To address these issues, numerous methods have been proposed (Fan and Xiang 2015; Cong et al. 2015; Gao et al. 2014; Bao et al. 2012; Wu et al. 2011; Fan et al. 2015; Possegger, Mauthner, and Bischof 2015; Mei and Ling 2011; Kwak et al. 2015; Hare, Saffari, and Torr 2011; Zhang, Ma, and Sclaroff 2014).

Recently, correlation filter has drawn increasing attention in computer vision due to its high efficiency and robust performance. Inspired by this, many correlation filter based trackers (Bolme et al. 2010; Danelljan et al. 2014a; Henriques et al. 2015; Ma et al. 2015; Zhu et al. 2016) are proposed and achieve good performance to some extend. Nevertheless, due to the sensitivity of correlation filter to occlusion, these trackers are prone to result in drift problem and even tracking failure in presence of occlusion. To deal with this problem, there are some attempts to apply part-based strategy to correlation filter for tracking (Liu, Wang, and Yang 2015; Li, Zhu, and Hoi 2015). This kind of trackers partition the object into multiple local parts, and each part corresponds with an independent correlation filter to estimate position. The final position of object is determined by combining the positions of all parts. Although these trackers can deal with occlusion in some degree, they still fail to track object in the situation of heavy occlusion or deformation because they ignore the spatial structure of object. (Liu et al. 2016) incorporates structure information by taking into account the motion model of each local part, however this method ignores the fact that the motion models of parts are constrained to that of global target and thus global appearance should be embedded into the model. In addition, it does not consider the temporal consistency of motion model which helps to alleviate the problem of drift away from object.

Inspired by these motivations, we in this paper propose a novel local-global correlation filter (LGCF) model for visual tracking. Our LGCF combines local and global appearance models by exploiting the circular shifts of global target and local parts for motion models to preserve the inner structure of object. For each local part, its motion model is represented by circular shifts and consistent with the motion model of global target. In this way, the spatial structure of target can be preserved. Due to self-deformation of object, however, the motion models of some parts may have slight discrepancies with that of global target. Therefore, we introduce the sparse constraint to model the relationship of motion models among global target and local parts, which helps our LGCF model tolerate outliers of local parts. To further improve the robustness of our model, we take into consideration temporal consistencies in both local parts and global target, and incorporate them into our model to alleviate the issue of drift away from object. In addition, to adapt our tracker to scale changes, we adopt an adaptive mechanism to estimate the scale of object, which is different from aforementioned part-based correlation filter trackers. Both qualitative and quantitative experiments on large-scale benchmark prove the ef-

fectiveness of our method.

In summary, we make the following contributions:

- We propose a novel LGCF model for visual tracking by exploiting the relationship of motion models among local parts and global target to preserve the structure of object.
- To reduce the risk of model drift, we introduce to incorporate temporal consistencies of both local parts and global target into our LGCF model to improve its robustness.
- Extensive experiments on OTB15 with 100 videos (Wu, Lim, and Yang 2015) demonstrate that our tracker performs favorably against state-of-the-art methods.

2. Related Work

Object tracking is one of the most challenging problems in computer vision and has been extensively studied (Wu, Lim, and Yang 2015; Pang and Ling 2013; Li et al. 2013). Here we would like to highlight two lines of works which are most relevant to ours.

The first line of works are to explore correlation filter for visual tracking. Owing to its high computational efficiency in Fourier domain, correlation filter has attracted extensive attention in object tracking. Bolme et al. (2010) propose a correlation filter tracker by learning a minimum output sum of squared error (MOSSE) for object appearance. Benefitting from the high computational efficiency of correlation filter, it has achieved real-time performance with a speed of several hundreds frames per second. Henriques et al. (2012) introduce kernel space into correlation filter and propose a circulant structure with kernel (CSK) method for tracking. Later in (Henriques et al. 2015), Henriques et al. further improve the performance of CSK tracker by replacing gray feature with HOG feature and propose the kernelized correlation filters (KCF) tracker. Danelljan et al. (2014b) present an adaptive correlation filter tracking algorithm by exploring color attributes of object. To solve the problem of scale changes, Danelljan et al. (2014a) propose a discriminative correlation filters based tracker (DSST) which learns multi-scale correlation filters to deal with the scale variations of object. Zhang et al. (2014) embed contextual information into correlation filters for tracking and Ma et al. (2015) propose a long-term correlation tracker with online random fern classifier. Despite promising results of aforementioned methods, their performances degrade in presence of occlusion. Different from above trackers, our method is able to handle occlusion by taking the benefits of part-based structure, and exploiting the relationship of motion models among local parts and global target to maintain structure of object further improves its robustness.

The second line of works are to exploit part-based strategy for visual tracking. One benefit of this strategy is its robust ability to resist occlusion. When target undergoes occlusion, remaining visible parts can still provide reliable information for tracking. Adam et al. (2006) propose to model target appearance with local patch histograms. Liu et al. (2013) utilize local sparse coding to model object appearance model, and Zhong et al. (2014) use global and local sparse representations to model target appearance and simply combine them for tracking. Nevertheless, due to high computational complexity of these methods, their performances are limited. To overcome this bottleneck, there are some attempts applying part-based strategy to correlation filter for visual tracking. Liu et al. (2015) propose to track target based on multiple object parts with multiple independent correlation filters. Li et al. (2015) suggest to find reliable patches to model object appearance and use these patches to estimate the state of target. The most similar work to ours is (Liu et al. 2016) where the structure of object is preserved by exploiting the relationship of local object parts with correlation filters. However, our method is significantly different from (Liu et al. 2016) in two aspects. First, our method takes into account the motion models of both local parts and global target and exploits their constrained relationship, while (Liu et al. 2016) only considers the relationship among local parts and ignores the importance of global target. Second, our tracker incorporates temporal consistencies of local parts and global target in our LGCF model to reduce the risk of model drift while (Liu et al. 2016) does not take any temporal information into consideration.

3. The Proposed Algorithm

3.1. Review of KCF Tracker

The KCF tracker (Henriques et al. 2015) models object appearance by a correlation filter $\mathbf{w}$ trained on an image patch $\mathbf{x}$ with $M \times N$ pixels, where all circular shifts of $\mathbf{x}_{m,n}$ are generated as training samples with Gaussian function label $\mathbf{y}_{m,n}$, $(m,n) \in \{0,1,\cdots,M-1\} \times \{0,1,\cdots,N-1\}$. The $\mathbf{w}$ can be derived through the following optimization

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{m,n} |\langle \phi(\mathbf{x}_{m,n}), \mathbf{w} \rangle - \mathbf{y}_{m,n}|^2 + \lambda \|\mathbf{w}\|^2 \quad (1)$$

where ϕ represents the nonlinear mapping function, and λ denotes the regularization parameter. Using fast Fourier transformation (FFT), Eq (1) is minimized as $\mathbf{w} = \sum_{m,n} \mathbf{a}(m,n)\phi(\mathbf{x}_{m,n})$, where the coefficient $\mathbf{a}$ is computed with

$$\mathbf{a} = \mathcal{F}^{-1}\big(\frac{\mathcal{F}(\mathbf{y})}{\mathcal{F}(\langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle) + \lambda}\big) \quad (2)$$

where $\mathbf{y} = \big\{\mathbf{y}_{m,n} | (m,n) \in \{0,1,\cdots,M-1\} \times \{0,1,\cdots,N-1\}\big\}$, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote Fourier transform and its inverse respectively. Given the learned $\mathbf{a}$ and appearance model $\hat{\mathbf{x}}$, we can compute the response map $\hat{\mathbf{y}}$ of a new patch $\mathbf{z}$ by

$$\hat{\mathbf{y}} = \mathcal{F}^{-1}\Big(\mathcal{F}(\mathbf{a}) \odot \mathcal{F}\big(\langle \phi(\mathbf{z}), \phi(\hat{\mathbf{x}}) \rangle\big)\Big) \quad (3)$$

where $\odot$ represents the Hadamard product. The target position is determined by the location of maximal value of $\hat{\mathbf{y}}$.

3.2. Local-Global Correlation Filter (LGCF) Model

First we transform Eq (1) to its equivalent dual form as the following

$$\min_{\mathbf{a}} \frac{1}{4\lambda} \mathbf{a}^{\mathrm{T}} \mathbf{X}\mathbf{X}^{\mathrm{T}} \mathbf{a} + \frac{1}{4} \mathbf{a}^{\mathrm{T}} \mathbf{a} - \mathbf{a}^{\mathrm{T}} \mathbf{y} \quad (4)$$

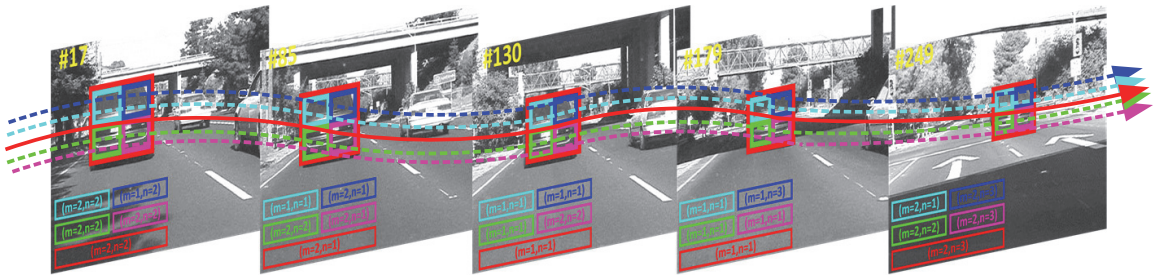


Figure 1: The five boxes at bottom-left corner of each frame denote the circular shifts of global target and individual parts. We can see that the shifts of object parts are close to the global object. Besides, we can also know that the temporal consistencies exist in both global target and individual object parts.

where $\mathbf{X} = [\mathbf{x}_{0,0}, \cdots, \mathbf{x}_{m,n}, \cdots, \mathbf{x}_{M-1,N-1}]^{\mathrm{T}}$, and $\mathbf{a}$ contains $M \times N$ dual optimization variables $\mathbf{a}_{m,n}$. Eq (1) and (4) are connected by $\mathbf{w} = \frac{1}{2\lambda}\mathbf{X}^{\mathrm{T}}\mathbf{a}$.

Global Model: The global model is trained on the entire object patch and can be expressed with

$$\min_{\mathbf{a}_g} \frac{1}{4\lambda}\mathbf{a}_g^{\mathrm{T}}\mathbf{H}_g\mathbf{a}_g + \frac{1}{4}\mathbf{a}_g^{\mathrm{T}}\mathbf{a}_g - \mathbf{a}_g^{\mathrm{T}}\mathbf{y}_g \tag{5}$$

where $\mathbf{H}_g = \mathbf{X}_g\mathbf{X}_g^{\mathrm{T}}$ and $\mathbf{X}_g$ represents the training samples of global target.

Local Model: Assume that the object is divided into K parts, and each part is corresponding with a correlation filter. Then our goal is to learn K weights $\{\mathbf{a}_k\}_{k=1}^K$ via the following optimization

$$\min_{\{\mathbf{a}_k\}_{k=1}^K} \sum_{k=1}^{K}(\frac{1}{4\lambda}\mathbf{a}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{a}_k + \frac{1}{4}\mathbf{a}_k^{\mathrm{T}}\mathbf{a}_k - \mathbf{a}_k^{\mathrm{T}}\mathbf{y}_k) \tag{6}$$

where $\mathbf{H}_k = \mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}$ $(k = 1, 2, \cdots, K)$ represents the training samples of the k^{th} localm part.

The basic idea of Eq (4) is to choose discriminative training samples $\mathbf{x}_{m,n}$ via $\mathbf{a}_{m,n}$ to distinguish the object from background. The training sample $\mathbf{x}_{m,n}$ represent all possible circular shifts which reflect the motion of object. Therefore, choosing training samples $\mathbf{x}_{m,n}$ via $\mathbf{a}_{m,n}$ is able to estimate the motion of target, which is to say the model model of target is encoded into $\mathbf{a}$.

In an ideal situation, the motion model of each local part should be consistent with that of global target. However, due to the self-deformation of object, the motion models of some local parts may be slightly different from that of global target. Thus we impose sparse constraint on the relationship of motion models among global target and local parts, which helps our LGCF model tolerate outliers of local parts. Figure 1 illustrates this thought. Based on this idea, we assume the relationship of motion models among global target and each local part as follows[1]

$$\mathbf{a}_k = \mathbf{a}_g + \boldsymbol{\delta}_k \tag{7}$$

[1]Note that when applying Eq (7), the $\mathbf{a}_g$ and $\mathbf{a}_k$ should be resized to have the same size. However, it does not influence the LGCF model.

where $\mathbf{a}_k$ and $\mathbf{a}_g$ represent motion models of local part k^{th} $(k = 1, 2, \cdots, K)$ and global target respectively. The $\boldsymbol{\delta}_k$ denotes the constraint between $\mathbf{a}_k$ and $\mathbf{a}_g$ and should be sparse.

Therefore, the goal of our LGCF model is to jointly learn the weights of global target $\mathbf{a}_g$ and local parts $\{\mathbf{a}_k\}_{k=1}^K$ in the following optimization

$$\begin{aligned}\min_{\mathbf{a}_g,\{\mathbf{a}_k\}_{k=1}^K} &\left\{ \sum_{k=1}^{K}(\frac{1}{4\lambda}\mathbf{a}_k^{\mathrm{T}}\mathbf{H}_k\mathbf{a}_k + \frac{1}{4}\mathbf{a}_k^{\mathrm{T}}\mathbf{a}_k - \mathbf{a}_k^{\mathrm{T}}\mathbf{y}_k) + \right. \\ &\left. \frac{1}{4\lambda}\mathbf{a}_g^{\mathrm{T}}\mathbf{H}_g\mathbf{a}_g + \frac{1}{4}\mathbf{a}_g^{\mathrm{T}}\mathbf{a}_g - \mathbf{a}_g^{\mathrm{T}}\mathbf{y}_g + \gamma\sum_{k=1}^{K}\|\boldsymbol{\delta}_k\|_1 \right\} \\ s.t. \quad &\mathbf{a}_k = \mathbf{a}_g + \boldsymbol{\delta}_k\end{aligned} \tag{8}$$

Between two consecutive frames, in fact, both objects and their contexts are similar. Thus the selected discriminative training samples should also be similar in consecutive frames. This is to say, the $\mathbf{a}$ in frame $(t-1)$ should be close to that in frame t, which is called *temporal consistency*. The *temporal consistency* exists in both global and local appearance models (see the colorful trajectories in Figure 2). Thus we revise our LGCF model in Eq (8) as follows

$$\begin{aligned}\min_{\mathbf{a}_g^t,\{\mathbf{a}_k^t\}_{k=1}^K} &\left\{ \sum_{k=1}^{K}(\frac{1}{4\lambda}(\mathbf{a}_k^t)^{\mathrm{T}}\mathbf{H}_k^t\mathbf{a}_k^t + \frac{1}{4}(\mathbf{a}_k^t)^{\mathrm{T}}\mathbf{a}_k^t - (\mathbf{a}_k^t)^{\mathrm{T}}\mathbf{y}_k^t) \right. \\ &+ \frac{1}{4\lambda}(\mathbf{a}_g^t)^{\mathrm{T}}\mathbf{H}_g^t\mathbf{a}_g^t + \frac{1}{4}(\mathbf{a}_g^t)^{\mathrm{T}}\mathbf{a}_g^t - (\mathbf{a}_g^t)^{\mathrm{T}}\mathbf{y}_g^t + \gamma\sum_{k=1}^{K}\|\boldsymbol{\delta}_k^t\|_1 \\ &\left. + \frac{\xi}{2}\|\mathbf{a}_g^t - \mathbf{a}_g^{t-1}\|^2 + \frac{\beta}{2}\sum_{k=1}^{K}\|\mathbf{a}_k^t - \mathbf{a}_k^{t-1}\|^2 \right\} \\ s.t. \quad &\mathbf{a}_k^t = \mathbf{a}_g^t + \boldsymbol{\delta}_k^t\end{aligned} \tag{9}$$

where $\frac{\xi}{2}\|\mathbf{a}_g^t - \mathbf{a}_g^{t-1}\|^2$ and $\frac{\beta}{2}\sum_{k=1}^{K}\|\mathbf{a}_k^t - \mathbf{a}_k^{t-1}\|^2$ model the temporal consistencies in global target and object parts, respectively.

3.3. Optimization

In this section, we use Alternating Direction Method of Multipliers (ADMM) method (Boyd et al. 2011) to solve the optimization in Eq (9). Using augmented Lagrange multipliers, we can incorporate the constraint condition into Eq (9) and

obtain the following Lagrangian function

$$
\begin{aligned}
&L(\mathbf{a}_g^t, \{\mathbf{a}_k^t, \boldsymbol{\delta}_k^t, \theta_k^t, \eta_k^t\}_{k=1}^K) = \\
&\Bigg\{ \sum_{k=1}^{K} \{ \frac{1}{4\lambda} (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{H}_k^t \mathbf{a}_k^t + \frac{1}{4} (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{a}_k^t - (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{y}_k^t \} + \\
&\frac{1}{4\lambda} (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{H}_g^t \mathbf{a}_g^t + \frac{1}{4} (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{a}_g^t - (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{y}_g^t + \gamma \sum_{k=1}^{K} \|\boldsymbol{\delta}_{l,k}^t\|_1 \\
&+ \frac{\xi}{2} \|\mathbf{a}_g^t - \mathbf{a}_g^{t-1}\|^2 + \frac{\beta}{2} \sum_{k=1}^{K} \|\mathbf{a}_k^t - \mathbf{a}_k^{t-1}\|^2 + \\
&\sum_{k=1}^{K} \{ (\theta_k^t)^{\mathrm{T}} (\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t) + \frac{\eta_k^t}{2} \|\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t\|^2 \} \Bigg\}
\end{aligned}
\tag{10}
$$

where θ_k^t and η_k^t are Lagrange multiplier and penalty parameter in frame t, and the new objective function becomes

$$
\min_{\mathbf{a}_g^t, \{\mathbf{a}_k^t, \boldsymbol{\delta}_k^t, \theta_k^t, \eta_k^t\}_{k=1}^K} L(\mathbf{a}_g^t, \{\mathbf{a}_k^t, \boldsymbol{\delta}_k^t, \theta_k^t, \eta_k^t\}_{k=1}^K) \tag{11}
$$

The ADMM algorithm iteratively updates one of the parameters by minimizing Eq (11) while keeping others fixed.
Update $\mathbf{a}_g^t$: Keeping other parameters fixed, $\mathbf{a}_g^t$ can be updated by solving Eq (12)

$$
\begin{aligned}
\mathbf{a}_g^t = \arg\min_{\mathbf{a}_g^t} \Big\{ &\frac{1}{4\lambda} (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{H}_g^t \mathbf{a}_g^t + \frac{1}{4} (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{a}_g^t - (\mathbf{a}_g^t)^{\mathrm{T}} \mathbf{y}_g^t \\
&+ \frac{\xi}{2} \|\mathbf{a}_g^t - \mathbf{a}_g^{t-1}\|^2 + \sum_{k=1}^{K} \{ (\theta_k^t)^{\mathrm{T}} \mathbf{a}_g^t + \frac{\eta_k^t}{2} \|\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t\|^2 \} \Big\}
\end{aligned}
\tag{12}
$$

and its solution is shown in Eq (13)

$$
\begin{aligned}
\mathbf{a}_g^t = &\Big(\frac{1}{2\lambda} \mathbf{H}_g^t + (\frac{1}{2} + \xi + \sum_{k=1}^{K} \eta_k^t) \mathbf{I} \Big)^{-1} \times \\
&\Big(\mathbf{y}_g^t + \xi \mathbf{a}_g^{t-1} + \sum_{k=1}^{K} (\theta_k^t + \eta_k^t \mathbf{a}_k^t - \eta_k^t \boldsymbol{\delta}_k^t) \Big)
\end{aligned}
\tag{13}
$$

where $\mathbf{I}$ denotes identity matrix.
Update $\boldsymbol{\delta}_k^t$: Keeping other parameters fixed, the problem in Eq (11) w.r.t to $\{\boldsymbol{\delta}_k^t\}_{k=1}^K$ can be decomposed into K independent sub-problems. The k^{th} sub-problem is

$$
\boldsymbol{\delta}_k^t = \arg\min_{\boldsymbol{\delta}_k^t} \Big\{ \gamma \|\boldsymbol{\delta}_k^t\|_1 - (\theta_k^t)^{\mathrm{T}} \boldsymbol{\delta}_k^t + \frac{\eta_k^t}{2} \|\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t\|^2 \Big\} \tag{14}
$$

The Eq (14) can be rearranged into the following problem according to (Boyd et al. 2011)

$$
\boldsymbol{\delta}_k^t = \arg\min_{\boldsymbol{\delta}_k^t} \Big\{ \frac{\gamma}{\eta_k^t} \|\boldsymbol{\delta}_k^t\|_1 + \frac{1}{2} \Big\| \boldsymbol{\delta}_k^t - (\mathbf{a}_k^t + \frac{\theta_k^t}{\eta_k^t} - \mathbf{a}_g^t) \Big\|^2 \Big\} \tag{15}
$$

and its solution can be obtained with

$$
\boldsymbol{\delta}_k^t = \mathcal{S}_{\frac{\gamma}{\eta_k^t}} (\mathbf{a}_k^t + \frac{\theta_k^t}{\eta_k^t} - \mathbf{a}_g^t) \tag{16}
$$

where $\mathcal{S}_\sigma(\mathbf{z}_i) = \mathrm{sign}(\mathbf{z}_i) \max(0, |\mathbf{z}_i| - \sigma)$ represents soft-thresholding operator for vector $\mathbf{z}$.
Update $\mathbf{a}_k^t$: Similarly, the problem in Eq (11) w.r.t $\{\mathbf{a}_k^t\}_{k=1}^K$ can also be decomposed into K sub-problems, and the k^{th} sub-problem is

$$
\begin{aligned}
\mathbf{a}_k^t = \arg\min_{\mathbf{a}_k^t} \Big\{ &\frac{1}{4\lambda} (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{H}_k^t \mathbf{a}_k^t + \frac{1}{4} (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{a}_k^t - (\mathbf{a}_k^t)^{\mathrm{T}} \mathbf{y}_k^t \\
&+ \frac{\beta}{2} \|\mathbf{a}_k^t - \mathbf{a}_k^{t-1}\|^2 + (\theta_k^t)^{\mathrm{T}} \mathbf{a}_k^t + \frac{\eta_k^t}{2} \|\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t\|^2 \Big\}
\end{aligned}
\tag{17}
$$

and its solution is shown as follows

$$
\begin{aligned}
\mathbf{a}_k^t = &\Big(\frac{1}{2\lambda} \mathbf{H}_k^t + (\frac{1}{2} + \beta + \eta_k^t) \mathbf{I} \Big)^{-1} \times \\
&\Big(\mathbf{y}_k^t + \beta \mathbf{a}_k^{t-1} - \theta_k^t + \eta_k^t (\mathbf{a}_g^t + \boldsymbol{\delta}_k^t) \Big)
\end{aligned}
\tag{18}
$$

Update θ_k^t and η_k^t: The Lagrange multiplier θ_k^t and penalty parameter η_k^t are updated as follows

$$
\theta_k^t = \theta_k^t + \eta_k^t (\mathbf{a}_k^t - \mathbf{a}_g^t - \boldsymbol{\delta}_k^t), \quad \eta_k^t = \tau \eta_k^t \tag{19}
$$

So far, we have introduced the solution for our LGCF model, as shown in Algorithm 1.

Algorithm 1 The solution for Eq (11)

Require: $\mathbf{H}_g^t$, $\mathbf{y}_g^t$, $\mathbf{H}_k^t$, $\mathbf{y}_k^t$, λ, γ, ξ, β, $\mathbf{a}_g^t$, $\mathbf{a}_g^{t-1}$, $\{\mathbf{a}_k^t, \mathbf{a}_k^{t-1}, \theta_k^t, \eta_k^t\}_k^K$;
while not converged **do**
  Update $\mathbf{a}_g^t$ based on Eq (13);
  for $k = 1$ to K **do**
    Update $\boldsymbol{\delta}_k^t$ based on Eq (16);
    Update $\mathbf{a}_k^t$ based on Eq (18);
    Update θ_k^t and η_k^t based on Eq (19);
  end for
end while
Return: Correlation filters $\mathbf{a}_g^t$, $\{\mathbf{a}_k^t\}_{k=1}^K$;

3.4. Tracking

After obtaining the global and local filters, we can estimate the positions of global target and each part. For global model, its response map in frame t can be computed with

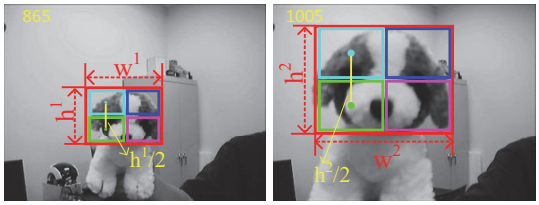
$$
\hat{\mathbf{y}}_g = \mathcal{F}^{-1} \Big(\mathcal{F}(\boldsymbol{\alpha}_g) \odot \mathcal{F}\big(\langle \phi(\mathbf{z}_g), \phi(\hat{\mathbf{x}}_g) \rangle \big) \Big) \tag{20}
$$

and the position $\bar{p}_g$ of global target is determined by the maximum value of $\hat{\mathbf{y}}_g$. For local model, the response map for the k^{th} part is calculated with

$$
\hat{\mathbf{y}}_k = \mathcal{F}^{-1} \Big(\mathcal{F}(\boldsymbol{\alpha}_k) \odot \mathcal{F}\big(\langle \phi(\mathbf{z}_k), \phi(\hat{\mathbf{x}}_k) \rangle \big) \Big) \tag{21}
$$

and the position p_k for part k is determined the maximum value of $\hat{\mathbf{y}}_k$. The object position is estimated using p_k via

$$
\bar{p}_k = p_k + \Delta_k, \quad k = 1, 2, \cdots, K \tag{22}
$$



(a) (b)

Figure 2: In image (a), the distance d^1 between indigo and green local parts is $h^1/2$. The ration of d^1 and object height is $1/2$. Likewise in image (b), the distance d^2 between indigo and green local parts is $h^2/2$, and the ration of d^2 and object height is $1/2$. Thus we can use this relationship to adaptively estimate the object scale.

where Δ_k denotes deformation vector (Zhang and Van Der Maaten 2014) between part k and the object center in last frame, and will be updated based on tracking result. The final position P of object is determined by the estimated positions of both global and local models as follows

$$P = w_g \bar{p}_g + \sum_{k=1}^{K} w_k \bar{p}_k \tag{23}$$

where w_g and $\{w_k\}_{k=1}^{K}$ represent the weights of global target and local parts, and are determined by the maximum values of their response maps as follows

$$w_g = \frac{f(\max(\hat{\mathbf{y}}_g))}{f(\max(\hat{\mathbf{y}}_g)) + \sum_{k=1}^{K} f(\max(\hat{\mathbf{y}}_k))} \tag{24}$$

$$w_k = \frac{f(\max(\hat{\mathbf{y}}_k))}{f(\max(\hat{\mathbf{y}}_g)) + \sum_{k=1}^{K} f(\max(\hat{\mathbf{y}}_k))} \tag{25}$$

where $f(z) = \frac{1}{1+\exp(-z)}$.

To address the problem of scale changes, we adopt an adaptive method by exploiting the relative distance among local parts as in (Akin et al. 2016). Although the scale of object changes all the time, the ration of relative distance among local parts and scale of object is stable. Figure 2 illustrates this idea. Thus the object scale S^t is estimated by

$$S^t = S^{t-1} \times \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\text{dist}(p_i^t, p_j^t)}{\text{dist}(p_i^{t-1}, p_j^{t-1})} \quad (i \neq j) \tag{26}$$

where $\text{dist}(\cdot,\cdot)$ denotes the Euclidean metric, and p_i^t represents the position of part i in frame t.

After obtaining the tracking result, we need to update both global and local models with current result. Different from (Henriques et al. 2015), we update the models based on their reliability. For global model, it is updated as follows

$$\boldsymbol{\alpha}_g^t = \begin{cases} (1-\epsilon)\boldsymbol{\alpha}_g^{t-1} + \epsilon \mathbf{a}_g^t, & f(\max(\hat{\mathbf{y}}_g)) > \Theta \\ \boldsymbol{\alpha}_g^{t-1}, & \text{otherwise} \end{cases} \tag{27}$$

$$\hat{\mathbf{x}}_g^t = \begin{cases} (1-\epsilon)\hat{\mathbf{x}}_g^{t-1} + \epsilon \mathbf{x}_g^t, & f(\max(\hat{\mathbf{y}}_g)) > \Theta \\ \hat{\mathbf{x}}_g^{t-1}, & \text{otherwise} \end{cases} \tag{28}$$

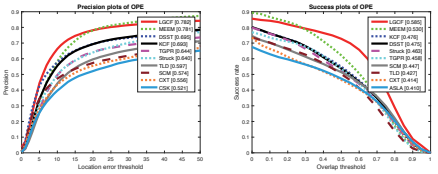


Figure 3: Comparisons of precision and success plots. Our LGCF tracker outperforms other state-of-the-art methods.

where Θ is a threshold. Similarly, the local model of part k is updated with

$$\boldsymbol{\alpha}_k^t = \begin{cases} (1-\epsilon)\boldsymbol{\alpha}_k^{t-1} + \epsilon \mathbf{a}_k^t, & f(\max(\hat{\mathbf{y}}_k)) > \Theta \\ \boldsymbol{\alpha}_k^{t-1}, & \text{otherwise} \end{cases} \tag{29}$$

$$\hat{\mathbf{x}}_k^t = \begin{cases} (1-\epsilon)\hat{\mathbf{x}}_k^{t-1} + \epsilon \mathbf{x}_k^t, & f(\max(\hat{\mathbf{y}}_k)) > \Theta \\ \hat{\mathbf{x}}_k^{t-1}, & \text{otherwise} \end{cases} \tag{30}$$

4. Experiments

Setting up: Our tracker is implemented in MATLAB on a 3.7 GHz Intel i7 Core PC with 12GB memory. The average running speed is around 8 frames per second. The γ, ξ and β are both set to 0.01. The Θ is ranging from 0.55 to 0.65. The number of local parts K is adaptively determined by the ratio of object size $\frac{O_w}{O_h}$, where O_w and O_h denotes object width and height, respectively. If $\frac{2}{3} \leq \frac{O_w}{O_h} \leq \frac{3}{2}$, target is divided into 2×2 local parts, i.e., $K = 4$; if $\frac{O_w}{O_h} < \frac{2}{3}$, target is divided into 3×1 local parts, i.e., $K = 3$; if $\frac{O_w}{O_h} > \frac{3}{2}$, target is divided into 1×3 local parts, i.e., $K = 3$. Other parameters are set to the same as the KCF tracker (Henriques et al. 2015). We use the same parameter values and initialization for all the sequences.

Dataset and evaluation metric: We evaluate the proposed algorithm on the OTB15 benchmark (Wu, Lim, and Yang 2015) with comparisons to 35 trackers including 31 trackers and other four recently published state-of-the-art trackers with their shared source codes: MEEM (Zhang, Ma, and Sclaroff 2014), TGPR (Gao et al. 2014), DSST (Danelljan et al. 2014a), KCF (Henriques et al. 2015). For better evaluation and analysis of our algorithms, the sequences are categorized according to 11 attributes, including scale variation, occlusion, deformation and so on. We employ the precision and success plots defined in (Wu, Lim, and Yang 2015) evaluate the robustness of the tracking algorithms.

Overall performance: Figure 3 shows the precision and success plots of our LGCF tracker and other methods. To make it clear, only the top 10 trackers are displayed. As shown in Figure 4, our proposed method ranks the first and achieves the best performance in both precision and success ranking plots. In specific, the proposed LGCF tracker achieves 0.782 ranking score in precision plots and 0.585 ranking score in success plots. Compared with the baseline KCF tracker with 0.693 precision ranking score and 0.476 success ranking score, our method obtains around 9% and 11% improvements, respectively. Besides, our tracker also

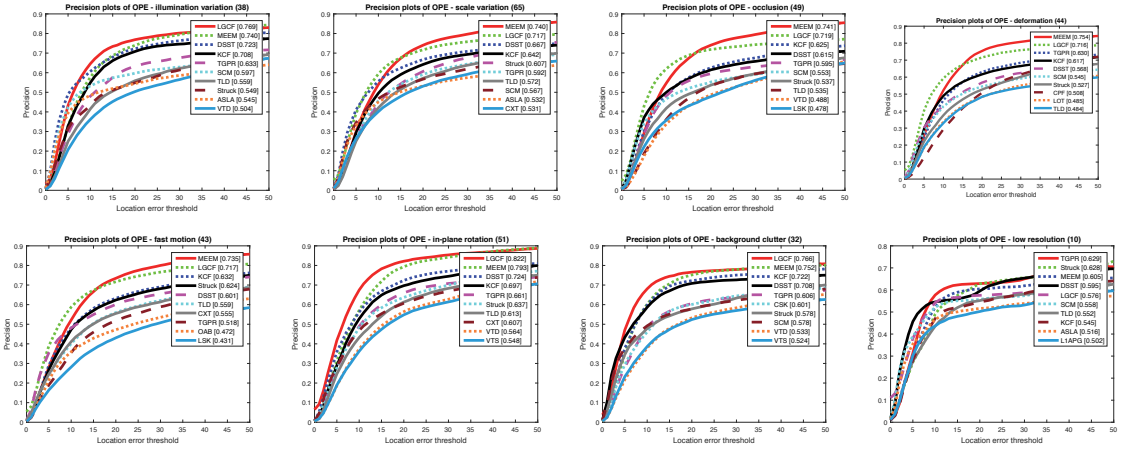


Figure 4: Comparisons of precision plots over eight tracking challenging of fast motion, background clutter, scale variation, deformation, illumination variation, occlusion, in-plane rotation and low resolution. Our proposed method performs favorably against other trackers.

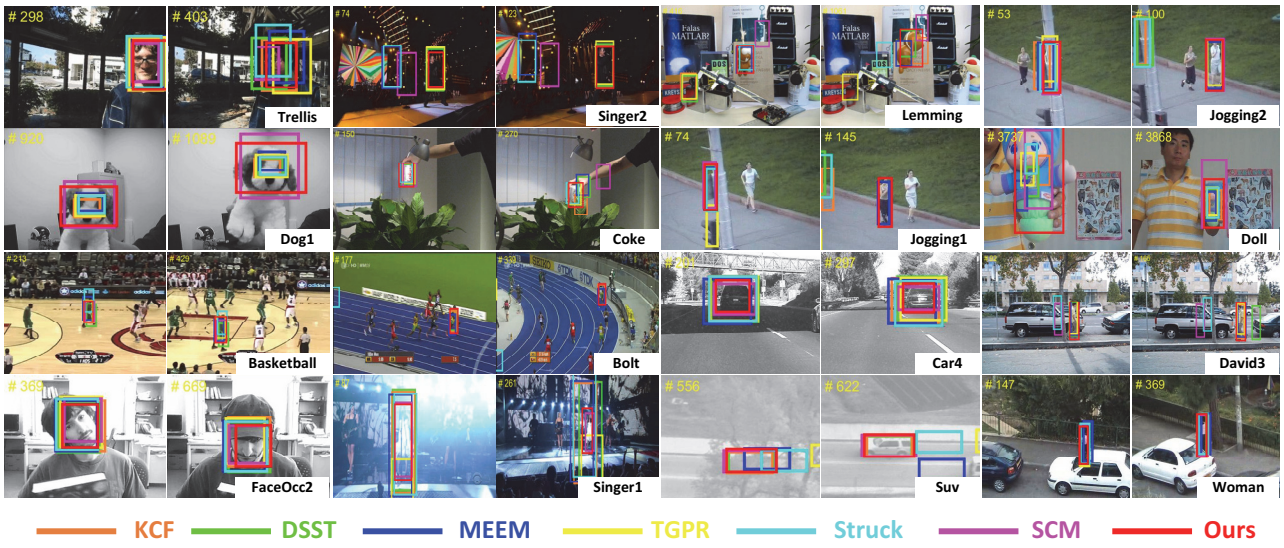


Figure 5: Qualitative results of seven trackers on eight sequences.

outperforms the state-of-the-art MEEM tracker with 0.781 precision ranking score and 0.53 success ranking score.

Attribute-based evaluation: The sequences in the benchmark dataset are annotated with 11 attributes to describe the different challenges in the tracking problem. These attributes are helpful for analyzing the performance of trackers in different situation. We report the performance of our tracker for eight challenging attributes in Figure 4. From Figure 4, we are able to know that the proposed LGCF achieves favorable results in seven attributes (within top 2), i.e., illumination variation, scale variation, occlusion, deformation, fast motion, in-plane rotation and background clutter. While in low resolution, our tracker only obtains the fifth ranking score.

Qualitative evaluation: We compare our tracker with six state-of-the-art methods: KCF (Henriques et al. 2015), MEEM (Zhang, Ma, and Sclaroff 2014), TGPR (Gao et al. 2014), DSST (Danelljan et al. 2014a), Struck (Hare, Saffari, and Torr 2011) and SCM (Zhong, Lu, and Yang 2014). The qualitative results are shown in Figure 5. From Figure 5, we can see that the proposed tracker performs well in illumination variations (*Trellis*, *Coke*, *Singer1* and *Singer2*), occlusion (*Lemming*, *Basketball*, *David3*, *Suv*, *FaceOcc2*, *Jogging1* and *Jogging2*), scale changes (*Dog1*, *Car4*, and *Doll*), deformation (*Woman* and *Bolt*), while other trackers can only handle some situations and degrade in other cases.

5. Conclusion

In this paper, we propose a novel LGCF model for visual tracking by exploiting the relationship of motion models among local object parts and global target to preserve the inner structure of target. In addition, to alleviate the issue

of model drift, we incorporate temporal consistencies of both local parts and global target in our model. Different from other trackers, we adopt an adaptive method to effectively estimate the scale of object. Extensive experiments on large-scale tracking benchmark with 100 image sequences demonstrate that the proposed algorithm performs favorably against state-of-the-art methods.

References

Adam, A.; Rivlin, E.; and Shimshoni, I. 2006. Robust fragments-based tracking using the integral histogram. In *CVPR*, 798–805.

Akin, O.; Erdem, E.; Erdem, A.; and Mikolajczyk, K. 2016. Deformable part-based tracking by coupled global and local correlation filters. *Journal of Visual Communication and Image Representation* 38:763–774.

Bao, C.; Wu, Y.; Ling, H.; and Ji, H. 2012. Real time robust l1 tracker using accelerated proximal gradient approach. In *CVPR*, 1830–1837.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. In *CVPR*, 2544–2550.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1):1–122.

Cong, Y.; Fan, B.; Liu, J.; Luo, J.; and Yu, H. 2015. Speeded up low-rank online metric learning for object tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 25(6):922–934.

Danelljan, M.; Häger, G.; Khan, F.; and Felsberg, M. 2014a. Accurate scale estimation for robust visual tracking. In *BMVC*.

Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; and Van de Weijer, J. 2014b. Adaptive color attributes for real-time visual tracking. In *CVPR*, 1090–1097.

Fan, H., and Xiang, J. 2015. Robust visual tracking with multitask joint dictionary learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Fan, H.; Xiang, J.; Liao, H.; and Du, X. 2015. Robust tracking based on local structural cell graph. *Journal of Visual Communication and Image Representation* 31:54–63.

Gao, J.; Ling, H.; Hu, W.; and Xing, J. 2014. Transfer learning based visual tracking with gaussian processes regression. In *ECCV*, 188–203.

Hare, S.; Saffari, A.; and Torr, P. H. 2011. Struck: Structured output tracking with kernels. In *2011 International Conference on Computer Vision*, 263–270.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 702–715.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):583–596.

Kwak, S.; Cho, M.; Laptev, I.; Ponce, J.; and Schmid, C. 2015. Unsupervised object discovery and tracking in video collections. In *ICCV*, 3173–3181.

Li, X.; Hu, W.; Shen, C.; Zhang, Z.; Dick, A.; and Hengel, A. V. D. 2013. A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology* 4(4):58.

Li, Y.; Zhu, J.; and Hoi, S. C. 2015. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *CVPR*, 353–361.

Liu, B.; Huang, J.; Kulikowski, C.; and Yang, L. 2013. Robust visual tracking using local sparse appearance model and k-selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(12):2968–2981.

Liu, S.; Zhang, T.; Cao, X.; and Xu, C. 2016. Structural correlation filter for robust visual tracking. In *CVPR*, 4312–4320.

Liu, T.; Wang, G.; and Yang, Q. 2015. Real-time part-based visual tracking via adaptive correlation filters. In *CVPR*, 4902–4912.

Ma, C.; Yang, X.; Zhang, C.; and Yang, M.-H. 2015. Long-term correlation tracking. In *ICCV*, 5388–5396.

Mei, X., and Ling, H. 2011. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(11):2259–2272.

Pang, Y., and Ling, H. 2013. Finding the best from the second bests-inhibiting subjective bias in evaluation of visual tracking algorithms. In *ICCV*, 2784–2791.

Possegger, H.; Mauthner, T.; and Bischof, H. 2015. In defense of color-based model-free tracking. In *CVPR*, 2113–2120.

Wu, Y.; Ling, H.; Yu, J.; Li, F.; Mei, X.; and Cheng, E. 2011. Blurred target tracking by blur-driven tracker. In *ICCV*, 1100–1107.

Wu, Y.; Lim, J.; and Yang, M.-H. 2015. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(9):1834–1848.

Zhang, L., and Van Der Maaten, L. 2014. Preserving structure in model-free tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(4):756–769.

Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; and Yang, M.-H. 2014. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, 127–141.

Zhang, J.; Ma, S.; and Sclaroff, S. 2014. Meem: robust tracking via multiple experts using entropy minimization. In *ECCV*, 188–203.

Zhong, W.; Lu, H.; and Yang, M.-H. 2014. Robust object tracking via sparse collaborative appearance model. *IEEE Transactions on Image Processing* 23(5):2356–2368.

Zhu, G.; Wang, J.; Wu, Y.; Zhang, X.; and Lu, H. 2016. Mc-hog correlation tracking with saliency proposal. In *AAAI*, 3690–3696.