Multi-Correlation Filters with Triangle-Structure Constraints for Object Tracking

Weijian Ruan, Jun Chen*, Member, IEEE, Yi Wu, Member, IEEE, Jinqiao Wang, Member, IEEE, Chao Liang, Ruimin Hu, Senior Member, IEEE, Junjun Jiang, Member, IEEE

Abstract—Correlation filters (CFs) have been extensively used in tracking tasks due to their high efficiency, although most of them regard the tracked target as a whole and are minimally effective in handling partial occlusion. In this study, we incorporate a part-based strategy into the framework of CFs and propose a novel multi-part correlation tracker with triangle-structure constraints. Specifically, we train multiple CFs for the global object and local parts, which are then jointly applied to obtain the correlation response of any candidate during tracking. The tracker is robust in handling partial occlusion because of the use of part-based representation. The remaining global representation can contribute reliable cues in cases where several local filters drift away in a specific scene. We further propose a triangle-structure model to measure the structural similarity of candidates. The model employs multiple triangles to determine the spatial relationship among parts and helps constrain the location of the target. Moreover, we introduce an effective part selection scheme based on energy and integrity, which is generally applicable to part-tracking models. Extensive experiments on two public benchmarks demonstrate the superiority of the proposed method over the state-of-the-art approaches.

Index Terms—Correlation filters, partial occlusions, triangle-structure, high-energy, high-integrity.

I. INTRODUCTION

Object tracking is one of the most challenging tasks in multimedia and computer vision, and it is related to a wide range of applications, such as multimedia analysis [1]–[4], video summarization and retrieval [5]–[7], and surveillance [8]–[10]. Although much progress has been achieved in recent years [11]–[13], the task remains challenging because of changes in target appearance caused by occlusions, deformation, illumination changes, scale variations, and other factors.

Most tracking algorithms are categorized as either generative or discriminative. Generative trackers perform tracking by searching the best-matching window. Various generative algorithms have been proposed, and these include sparse representation [14]–[17], incremental subspace learning [18], and density estimation [19]. Meanwhile, discriminative methods aim to build a model to distinguish the target from backgrounds.

These methods, typically learn classifiers through P-N learning [20], multiple-instance learning [21], structured-output SVMs [22], and correlation filters (CFs) [23], [24]. Overall, extensive experiments conducted on tracking benchmarks have shown that discriminative methods are more competitive than generative approaches in most testing scenarios [25]. We focus on developing a discriminative tracker in this work.

CF trackers can be categorized as discriminative methods. These trackers elicited much attention in object tracking due to their computational efficiency and competitive performance [23], [24], [26]–[29]. CF is designed to produce a correlation peak for the target in a scene while yielding a low response to the background, and they are usually employed to detect expected patterns. The intrinsic idea of correlation tracking is to track the target by correlating the filter over a large search window in the subsequent frame, and the location with the maximum correlation response indicates the new location of the target. Specifically, the correlation can be calculated in the Fourier domain to avoid the time-consuming convolution operation, because the correlation in the time domain corresponds to an element-wise multiplication in the Fourier domain, as proven by the convolution theorem. However, most existing CF trackers regard the target as a whole and train a holistic filter for tracking. Therefore, these trackers show reduced robustness and effectiveness when dealing with partial occlusions due to the disregard of local information, as shown in the top row of Fig. 1. For example, the two recent correlation trackers, DSST [27] and KCF [26], have achieved state-of-the-art results and...
outperform all other attended trackers in terms of accuracy in VOT challenge [30], but they tend to drift away when partial occlusion occurs.

Several methods have been proposed recently to address the partial occlusion problem by applying the part-based tracking strategy [31]–[38]. These methods regard the target as a collection of local patches and adopt part-based representation to construct an observation model. Even when the target is partially occluded, the visible parts still provide reliable cues for capturing the target, as shown in the bottom row of Fig. 1. The tracking object was divided into small patches by using a regular grid in [31], and $l_1$ sparsity was adopted as similarity metric to locate the closest candidates in the subsequent frame. RPT [37] adopts a trackable confidence function to compute and select the reliable patches, and tracks the target by identifying and exploiting the most reliable patches. However, existing part-based trackers still exhibit problems. First, these trackers only adopt local representation to model the target and disregard the holistic appearance, which may contribute adversely to situations without partial occlusions. Second, these methods independently regard each part and neglect the spatial relationship among parts. Third, they select parts with a regular grid and usually fail to select discriminative parts to track. In addition, an efficiency limitation is commonly observed when the part-based tracking strategy is applied because the target is divided into multiple patches to track in single frame instead of tracking the target itself [33], [36].

To this end, we incorporate the part-based strategy into the CF framework to model the target’s appearance and propose a novel multi-part correlation tracker with triangle-structure constraints (MCTTC). By exploiting the global and local appearances of the target to construct multiple CFs, the target location is estimated using the filters to jointly measure the correlation responses of the holistic candidate and its local parts within a search window over time, as shown in Fig. 2. Specifically, the tracker becomes robust to partial occlusion due to the metric of part-based representation. Meanwhile, preserving the holistic presentation contributes remarkably when local part filters drift away in a specific scene (e.g., low resolution). The computational cost of our tracker is greatly reduced due to the high efficiency of CF. Furthermore, in comparison with prior studies that independently processed parts, our study proposes a triangle-structure model to describe the spatial relationship among parts and provides a structural similarity to help constrain the target location. A triangle is constructed for every three parts by connecting their centers initially. Thus the structure of the target is determined by several triangles. In consequent frames, the structural similarity of any candidate is measured by estimating how similar the triangles formed in candidate are to the ones formed in the original tracking target. The process of selecting discriminative parts is important for part-based models. Hence, we introduce an effective part selection scheme by cropping out the local parts with high energy and integrity via superpixel division and exemplar-SVM [39], which is generally applicable to part-based tracking models. Extensive experiments conducted on two public benchmarks demonstrate that the proposed method outperforms the state-of-the-art methods.

Generally, the main contributions of our work can be summarized as follows:

- We propose a novel multi-part correlation tracker that tracks the target by integrating the correlation responses of the holistic candidate and its local parts. The method can effectively handle partial occlusions.
- We propose a spatial triangle-structure model to enhance location precision. This method employs several triangles to describe the structure of the target and track the target location.
- We propose an effective discriminative part selection strategy based on energy and integrity. The proposed strategy is generally applicable to part tracking models.

II. RELATED WORK

Object tracking has been studied extensively in recent decades [25], [40]. In this section, we introduce the two categories of tracking methods, namely, generative and discriminative trackers. Correlation and part-based trackers used in previous related studies are also presented.

A. Generative Methods

In generative methods, tracking is formulated as searching for regions that are the most similar to the target object within a neighborhood. Examples of generative methods include VTD [41], L1 [42], LSHT [43], SST [15], and SPC [44]. VTD [41] effectively extends the conventional particle filter framework with multiple motion and observation models to account for appearance variations. The L1 tracker [42] represents candidates sparsely by using $l_1$ norm minimization. LSHT [43] adopts a locality-sensitive histogram that exploits the spatial weight of every pixel. SST [15] exploits the intrinsic relationship among target candidates and their local patches while preserving the spatial layout structure to determine sparse representations jointly. SPC [44] is a context-based tracking method that employs a new saliency prior context model, which can track objects by sequentially maximizing the computed confidence map of the target location in each frame. Generally, despite the efficiency limitation of these trackers, they are robust to object occlusion but sensitive to similar distracters in the surrounding area of the target object.

B. Discriminative Methods

Discriminative methods regard the tracking process as a binary classification task that aims to determine the decision boundary for separating the target object from the background. Our proposed method is a discriminative tracker. We introduce several classic discriminative methods, namely, TLD [20], AOGTracker [45], Struck [22], and MVSVM [46]. In [20], Kalal et al. proposed the P-N learning algorithm to exploit the underlying structure of positive and negative samples and determine classifiers for object tracking. Struck [22] uses structure output learning for visual tracking to avoid the label prediction problem in common online classifiers and has exhibited good performance. In [47], Chu et al. utilized a projected gradient to assist multiple kernels in finding
the best match during tracking under predefined constraints. AOOGTracker [45] simultaneously combines tracking, learning, and parsing objects with hierarchical and compositional And-Or graph (AOOG) representation to handle occlusion and background clutter. MVSVM [46] adopts a multi-view learning framework using multiple SVMs based on multiple features. In most situations, discriminative trackers are more effective than generative trackers [11], [25].

### C. Correlation Tracking

The use of CFs in tracking tasks has elicited considerable attention recently due to the robustness and computational efficiency of CFs [23], [26]–[29], [48]. MOSSE [23] is a fast correlation tracker with a minimum output sum of squared error filter that can handle hundreds of frames per second. STC [48] exploits the spatiotemporal context in a Bayesian framework for visual tracking. CSK [49] is used to train a filter based on kernel ridge regression and can formulate kernelized CFs by using circulant matrices to produce thousands of samples for training and prediction. CSK has been improved by incorporating multi-channel features in the Fourier domain, namely, KCF [26]. Several variations of KCF trackers have been investigated to improve tracking performance. For example, DSST [27] learns separate filters for translation at different scales to adapt to the scale variation, and MUSTer [28] employs short- and long-term memory stores inspired by a psychological memory model. Staple [24] combines two image patch representations that are robust to color changes and deformations by exploiting the advantages of color statistics and CF. Although remarkable progress has been made in correlation trackers, most of these trackers tend to model the target as a whole and cannot handle partial occlusions well.

### D. Part-based Trackers

Many methods have been proposed to deal with partial occlusion, and remarkable progress has been achieved by applying a part-based strategy [16], [32], [34]–[37], [50]. In HAB [50], objects are tracked by matching the intensity histograms and updating the shape and by adjusting small blocks of the tracked window. In [32], the part-based tracking problem was addressed by using latent structured learning. PMT [34] tracks an object by matching the parts among the frames. SCM [16] uses a sparsity-based collaborative model to exploit holistic templates and local representations. In [51], a flock of trackers was proposed to track a set of patches with a regular grid structure that allows certain drifts. RPT [37] addresses the tracking problem by exploiting and identifying the reliable patches of the object. In BLC [36], boosted local classifiers are trained based on local patches with multiple SVMs to address the tracking problem. Notably, RPACF [38] appears to be similar to our proposed model; it also adopts part filters during tracking but only considers local information and does not model the target’s structure. Overall, these recent proposals confirm that the part-based approach can address the occlusion problem, but the problem is still far from being solved due to the issues discussed in Section I.

### III. Proposed Approach

Fig. 2 shows the tracking framework of the proposed MCTTC. First, we initialize multiple tracking parts, including one holistic part (target itself) and several local parts with high energy and high integrity, via superpixel division and exemplar-SVM. Second, we train a specific CF for each of the parts. During tracking, part filters are applied to obtain the correlation response for the input image. After combining them with different weights, we adopt a particle filter framework to sample candidates on the correlation response map. Finally, we measure the confidence of a candidate by the correlation response of all parts in the candidate and structural similarity, which is described by the spatial triangle-structure model presented in Section III (B).

#### A. Multi-part Correlation Tracker

1) **Part Correlation Tracking:** CF trackers have achieved impressive results on public tracking benchmarks [11], which model the object appearance by using a CF $w$ trained on an image patch $x$ of size $M \times N$, where all the circular shifts of $x_{m,n}$, $(m,n) \in \{0,1,...,M-1\} \times \{0,1,...,N-1\}$ are generated as training samples with the Gaussian function label $y_{m,n}$. The goal is to find a function $f(z) = w^Tz$ that minimizes the squared error over samples $x_{m,n}$ and their regression targets $y_{m,n}$, that is, to find the optimal weights $w$ via the following optimization problem.

$$w = \arg \min_{w} \sum_{(m,n)} |\langle \phi(x_{m,n}), w \rangle - y_{m,n}|^2 + \lambda \|w\|^2,$$

(1)

where $\phi$ denotes the mapping to a Hilbert space induced by the kernel $\kappa$, defining the inner product as $\langle \phi(x), \phi(x') \rangle = \kappa(x,x')$. A Gaussian kernel $\kappa(x,x') = \exp(-|x-x'|^2/\sigma^2)$ is adopted as the mapping function. The constant $\lambda$ is a regularization parameter. By using fast Fourier transformation to compute the correlation, the weight in Eq. (1) is minimized as $w = \sum_{m,n} \alpha(m,n) \phi(x_{m,n})$, and the coefficient $\alpha$ is calculated as

$$\alpha = F^{-1} \left( \frac{F(y)}{F(\langle \phi(x), \phi(x) \rangle) + \lambda} \right),$$

(2)

where $y = \{y_{m,n}\}$. $F$ and $F^{-1}$ denote the Fourier transform and its inverse, respectively. Given the learned $\alpha$ and target appearance model $\hat{x}$, the tracking task is performed on an image patch $z$ in the new frame with the search window size $M \times N$ by computing the response map as follows:

$$\hat{f}(z) = F^{-1} (F(\alpha) \odot F(\langle \phi(z), \phi(\hat{x}) \rangle)),$$

(3)

where $\odot$ is the element-wise product. The higher the $\hat{f}(z)$ of the position is, the more likely the target occurs.

Different from the prior researchers, we incorporate the part-based strategy with CFs to address the tracking problem, especially in situations with partial occlusions and deformation. Initially, we apply Eq. (2) to train multiple CFs $\{\alpha_i\}$ for holistic and local parts, thus fully exploiting the structural information of the tracking object. For each selected part, we adopt its appearance and the surrounding context centered on the part to train a specific part filter, as shown in Fig. 3. In
consecutive frames, the part CFs work and can obtain several sub-response maps \( \{ f_i^t \} \) for the input image by using Eq. (3) at time \( t \), which are then combined into a holistic response map \( R_t \) and can be formulated as

\[
R_t = \sum_{i=1}^{K} \beta_i^t (H_i \circ f_i^t),
\]

(4)

\[
h_{jk} = \exp \left( -2 \cdot \frac{|z|^2}{\delta^2} \right), \quad h_{jk} \in H_i,
\]

(5)

where \( K \) is the number of parts; \( H_i \) is the post-processed window, which can decrease the response noise from other sub-responses because a partial overlap occurs among the parts; \( z \) is the distance between the pixel and the position with maximum response in \( f_i^t \); and \( \delta^2 \) is the size of \( f_i^t \). In addition, we assign different weights \( \beta_i^t \) to each part filter according to the Peak-to-Sidelobe Ratio (PSR) value of the corresponding part, which is an appropriate index to evaluate the quality of tracking. To compute PSR, the correlation output (response) is split between the peak (i.e., the maximum response value) and the sidelobe (i.e., the rest of pixels around the peak), which is expressed as follows:

\[
\beta_i = \text{PSR}_i = \frac{\max (f_i^t) - \mu_i}{\sigma_i},
\]

(6)

where \( \mu_i \) and \( \sigma_i \) denote the mean and standard deviation of the sidelobe, respectively.

2) Tracking Candidates with Particle Filters: During the tracking process, we draw \( \Omega \) candidates \( \{ s_i \}_{j=1,2,\ldots,\Omega} \) with the framework particle filter [52], which can effectively estimate and propagate the posterior probability density function of state variables regardless of the underlying distribution. Given the observation set of target \( z_1:t = \{ z_1, z_2, \ldots, z_t \} \) up to time \( t \), the target state variable \( s_t \) can be computed by maximizing a posteriori estimation as follows:

\[
\hat{s}_t = \arg \max_{s_i^t} P(s_i^t | z_1:t),
\]

(7)

where \( s_i^t \) indicates the state of the \( i \)-th candidate. The posterior probability \( P(s_t | z_1:t) \) can be inferred with the Bayesian theorem recursively as follows:

\[
p(s_t | z_1:t) \propto p(z_t | s_t) \int P(s_t | s_{t-1})p(s_{t-1} | z_{1:t-1})ds_{t-1},
\]

(8)

where \( p(s_t | s_{t-1}) \) denotes the dynamic model that describes the temporal correlation of the target states between consecutive frames. \( p(z_t | s_t) \) denotes the observation model that denotes the likelihood of the observation \( z_t \) at state \( s_t \). It plays an important role in robust tracking. In our method, we measure this by the correlation response and structural similarity for \( s_t \), i.e., for all the candidates.

We sample candidates based on the tracking result of the previous frame during tracking. Similar to [16], we employ affine transformation with six parameters to model the object motion between two consecutive frames. The state transition is formulated as

\[
p(s_t | s_{t-1}) = N(s_t; s_{t-1}, \Sigma),
\]

(9)

where \( N \) denotes a Gaussian distribution, and \( \Sigma \) is a diagonal covariance matrix whose elements are the variances of the affine parameters.

However, the variances of six affine parameters are not fixed for all videos in our proposed method but are adaptive based on the statistical variations of object motion in the previous several frames of a video. For example, for the scenario...
with fast motion, the resolution should be high to expand the sampling region. Specifically, we adaptively change the sampling range by the average fluctuation degree of object motion in the previous 20 frames of any tested video, which is measured based on the tracking results in these frames.

As shown in Fig. 4, we sample \( A \) candidates around the tracking result in the previous frame after obtaining the holistic response map of the current frame. For each candidate, we crop out \( K \) parts according to the relative ratios. Then, we measure its correlation response \( p_r \) via part center pooling and its structural similarity \( p_{tri} \) based on max pooling. Thus, the confidence of candidate \( s_j \) is defined as

\[
p(s_j) = \begin{cases} 
(1 - \mu)p_r + \mu p_{tri}, & \text{if } p_r > p_{tri} \\
(p_r + p_{tri})/2, & \text{otherwise}
\end{cases},
\]

(10)

where \( \mu = p_{tri}/(2p_r) \). In this manner, \( p_r \) is given a higher weight than \( p_{tri} \) as a form of correction in cases where the wrong structural constraint is introduced when undergoing remarkable changes in appearance. \( p_{tri} \) denotes the structural similarity of the candidate described using the spatial triangle-structure model, which will be presented in detail in the subsequent section. We compute \( p_r \) by summing the correlation responses of the part centers, that is, center pooling (Fig. 4).

\[
p_r = \frac{1}{K} \sum_{i=1}^{K} R_{tri}^i (l_c^i),
\]

(11)

where \( l_c^i \) denotes the center of the \( i \)-th part in the candidate.

Therefore, the target can be estimated as the candidate with maximum confidence \( p(s_j) \) in Eq. (10).

**B. Spatial Triangle-Structure Model**

Considering the structural stability of a triangle, we adopt several triangles to describe the spatial relationship among parts to model the object structure. Thus, the structural similarity of any candidate can be estimated by measuring how similar the new triangles formed in the candidate are to the original ones.

1) **Constructing Positive Triangle Groups:** In the first frame, we connect the part centers to initialize the structure of the target, and the centers of all the part tracking results in the previous \( T \) frames (\( T = 20 \) in our experiments) on the basis of the argument that part CFs can accurately estimate the parts in the former \( T \) frames of a video. In this manner, we can capture several positive structures of the tracking object, as shown in the top row of Fig. 5.

In line with the theory of permutation and combination, for \( K \) parts in our framework, \( Q = K(K-1)(K-2)/6 \) triangles can be constructed with every three parts as one. Therefore, we can obtain \( Q \) triangles from the structure of the object in each frame. With the constructed positive structures in the consequent \( T \) frames, several positive triangle groups \( \{G_i\}_{i=1,2,...,Q} \) are generated, with each one possessing \( T \) triangles. Thus, a positive triangle group describing the spatial relationship of any three parts exists.

In particular, we map the triangle into the space rectangular coordinate system with the length of the three sides of the triangle as the three directional coordinates. Thus, one triangle becomes a point in the coordinate system. Then, any triangle group \( G_i \) becomes a set of points and we represent it by an ellipse to include all positive triangle samples, as shown in the middle row of Fig. 5. During the tracking process, the structural similarity of the candidate can be measured based on the distances (\( d \)) between the estimated triangles and the ellipses’ centers.

\[
a^2 (\ell_x - \ell_{x_0})^2 + b^2 (\ell_y - \ell_{y_0})^2 + c^2 (\ell_z - \ell_{z_0})^2 = 1,
\]

(12)

\[
a = 2/\left( \max\{\ell_{x_0}\}_{j=1,2,...,T} - \min\{\ell_{x_0}\}_{j=1,2,...,T} \right).
\]

(13)

\[
b = 2/\left( \max\{\ell_{y_0}\}_{j=1,2,...,T} - \min\{\ell_{y_0}\}_{j=1,2,...,T} \right).
\]

(14)

\[
c = 2/\left( \max\{\ell_{z_0}\}_{j=1,2,...,T} - \min\{\ell_{z_0}\}_{j=1,2,...,T} \right).
\]

(15)

2) **Measuring Structural Similarity:** As shown in the bottom row of Fig. 5, we initially crop out \( K \) parts with the same ratios as that when initializing the parts. Then, we employ max pooling to select the positions with the maximum correlation response in the \( 5 \times 5 \) window centered on the parts. By
connecting them to describe the structure of the candidate, we obtain \( Q \) estimated triangles. Thus, the structural similarity of the candidate can be formulated as

\[
p_{tri} = \frac{1}{Q} \sum_{i=1}^{Q} \exp(-\gamma d_i),
\]

where \( d_i \) denotes the distance between the \( i \)-th triangle and the center of \( G_i \), namely

\[
d_i^2 = (x_i - x_{i1})^2 + (y_i - y_{i1})^2 + (\ell_z i - \ell_z i1)^2.
\]

\( \gamma = 20 \) is a constant. Notably, although a part may drift away slightly from its accurate position due to rotation, scale changes, deformation, and other factors, the hypothesis remains. The position with maximum response in the part should always converge to the real position (the peak) because the distribution of the response is similar to a Gaussian distribution. Hence the structure should be preserved to some extent.

C. Model Update

In object tracking, adaptively updating the model is important because the object appearance may change due to illumination, deformation, occlusions, and other factors. The model in most CF trackers consists of the learned target appearance and the transformed filter coefficients. They are computed only by considering the current appearance. The tracker then updates the filter coefficients by simple linear interpolation with a fixed learning rate. Thus the appearance and CF are updated without any constraints. Once the tracker drifts away, the model is damaged in the remaining frames. We address this problem by using an adaptive updating scheme. In our method, an occluded part should not be updated to avoid introducing the wrong appearance. Therefore, we set a threshold \( \rho \) to adaptively update each part and determine whether the tracking result is reliable or not. The parts with lower detection scores should be minimally updated because they seem to drift away. The updating scheme is defined as

\[
F(\alpha)_i^t = \begin{cases} 
(1 - \eta) F(\alpha)_{i-1}^t + \eta F(\alpha), & \text{if } \beta_i^t > \rho, \\
F(\alpha)_{i-1}^t, & \text{otherwise},
\end{cases}
\]

\[
x_i^t = \begin{cases} 
(1 - \eta) \hat{x}_i^{t-1} + \eta \hat{x}^t, & \text{if } \beta_i^t > \rho, \\
\hat{x}_i^{t-1}, & \text{otherwise},
\end{cases}
\]

where \( \eta \) is the learning rate and fixed to 0.02. Notably, when the target is tracked correctly, \( PSR \) typically ranges between 6 and 13, which indicates a very strong peak. When \( PSR \) decreases to around 4, the detection of the corresponding part is unreliable. Thus, we set \( \rho \) to 3.

D. Selection of Discriminative Parts

How to select proper parts is crucial in part-based tracking methods. To fully exploit the structural information of the target, we select a holistic part (the target itself) and multiple local parts with high energy and high integrity.

1) Energy: Inspired by deformable part-based models [53], we select the intra-parts that cover high-energy regions of the root filter, where the energy of a region is defined by the norm of the positive weights in the region. We use exemplar-SVM [39] to measure the energy of parts. The key idea of exemplar-SVM is that, for each positive sample (also regarded as an exemplar), an exemplar-SVM is trained by the corresponding set of samples, where only one positive sample exists and the rest are negative. We randomly use \( N_E \) negative samples with the regions \( r \geq 50 \) (\( r \) is the distance between their centers and the positive sample). Thus, the object function of exemplar-SVM \( f_E(x) \) can be expressed as

\[
f_E(x) = w_E^T x + b_E,
\]

where the weight \( w_E \) and the offset \( b_E \) can be solved by optimizing the convex objective function as follows:

\[
\Omega(w, b) = ||w||^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in E_{\neg E}} h(-w^T x_E - b),
\]

where \( x_E \) denotes the positive sample, and the hinge loss function is used in \( h(x) = \max(0, 1 - x) \). \( C_1 \) and \( C_2 \) are regularization parameters. We can obtain the weight matrix after training and evaluate the energy \( E \) of each region by:

\[
E = \sum_{(x, y) \in B_i} \max(0, w_E(x, y), b_E).
\]

2) Integrity: The integrity of a part is determined by how likely a part can be treated as an integral part. To measure integrity, we apply the SLIC algorithm [54] to extract superpixels of the object. The spatial proximity weight and number of superpixels are set empirically (10 and 10, respectively, in our experiments). Thereafter, the integrity \( w_i^t \) of the \( i \)-th part \( B_i \) is defined as follows:

\[
w_i^t = \max \left| \frac{S_i^t}{|B_i|} \right| \delta_j,
\]

where \( S_i^t \) denotes the \( j \)-th superpixel within part \( B_i \). In addition, \( \delta_j = \exp(-2d_j^2/\sigma^2) \), where \( d_j \) is the distance between the centers of \( S_i^t \) and \( B_i \), which can be estimated based on the minimum enclosing rectangle of \( S_i^t \). \( \sigma \) is the scale parameter of \( B_i \).

We select the parts with high energy and high integrity and express the selection as follows:

\[
B^* = \arg \max_{B_i \subset B} \sum_{(x, y) \in B_i} \max(0, w_i^t w_E(x, y)).
\]

On the basis of Eq. (23), we can select the local part with the highest integrity and energy. Once a part is selected, the corresponding pixels are set to 0 for the next selection to avoid the excessive overlapping of the local parts. In this way, we can select multiple local parts aside from the holistic part that includes the entire target.
IV. IMPLEMENTATION

To demonstrate the proposed tracker, some technical details and a complexity analysis are provided below.

**Feature extraction.** We represent each feature vector by a concatenation of multi-channel HOG features with 31 bins and cells of $4 \times 4$, such as in [26]. However, our tracker is generic, and any feature with preserved spatial information can be incorporated.

**Triangle-structure model.** When selecting positive triangle samples to construct the triangle group, we use a threshold $\rho_1$ to constrain the selection. That is, only when $\left(PSR_i \times \max \left(\hat{f}_i\right)\right) > \rho_1$ will the corresponding part be adopted to construct the estimated triangles. Considering that the target may encounter scale variations in consequent frames, we normalize all the triangles, such as $[\ell_x, \ell_y, \ell_z] = [\ell_x, \ell_y, \ell_z]/\max \{\ell_x, \ell_y, \ell_z\}$.

**Computational complexity analysis.** During tracking, filter training is just conducted on an image patch for each part frame by frame, so the corresponding complexity is $O(K)$. Given that the target estimation involves obtaining the correlation response and structural similarity, the corresponding complexity is $O(AK) + O(AQ)$. Hence, the total complexity of our proposed tracker is $O(K + AK + A)$. As indicated in Eq. (2,3), the full kernel correlation can be efficiently computed in the Fourier domain because only a few elementary operations exist. Therefore, a minimal amount of time is required to train filters and estimate correlation responses, i.e., the two components, $O(K)$ and $O(AK)$ can be efficiently addressed, which remarkably mitigates the computational cost of our tracker. Furthermore, only a few simple operations for the third one $O(AQ)$ are needed, and only slight effects on the computation are observed.

V. EXPERIMENTS

All our experiments are performed using MATLAB R2015b on a PC with a 2.8 GHZ Intel Core i5 processor and 4GB of RAM. The proposed tracker runs a speed of about 10 frames per second (fps).

A. Experimental Setup

**Datasets.** To evaluate the performance of the proposed tracker, we conduct extensive experiments on two datasets, namely, OTB50 [11] and OTB100 [25], which involve 50 and 100 image sequences, respectively. The videos have several scenarios (e.g., surveillance, entertainment, and sports). The image sequences are tagged with 11 attributes, which represent the challenging aspects in object tracking (e.g., occlusions, scale variation, deformation, etc.). Thus, each attribute covers several image sequences whose number are shown in Table I.

**Parameter settings.** In the evaluations, we find that when the size of intra-parts is set between 0.5 and 0.6 of the target size, the tracker is relatively effective. Thus, we fix the size of the intra-parts to 0.5 width and 0.6 height of the target size. The number of parts is set to $K = 5$ (i.e., one holistic and four local parts); thus, $Q = K(K-1)(K-2)/6 = 10$. Moreover, $A = 400$ candidates are sampled in each frame during tracking. The threshold $\rho_1$ for constraining the positive triangle samples is set to 2.6.

**Comparison of state-of-the-art trackers.** We conduct comparisons with 15 existing state-of-the-art trackers, namely, (i) convolutional neural network based tracker (CNT [55]); (ii) CF trackers (Staple [24], KCF [26], DSST [27], STC [48], and CSK [49]); (iii) part-based trackers (RPT [37], SCM [16], and IVT [18]); (iv) structured SVMs (Struck [22]); and (v) other classical trackers (LOT [56], TGPR [57], TLD [20], VTD [41], and MIL [21]). Notably, Struck and SCM are the top two trackers in the original evaluations on OTB50.

For fair comparisons, we evaluate the proposed MCTTC against the above state-of-the-art methods using publicly available source codes or the tracking result files provided by the researchers. The default parameters are set for initialization. For each sequence, the location of the target is manually labeled in the first frame.

B. Quantitative Evaluation

To quantitatively evaluate the performance of each tracker, we use the center location error (CLE) and the overlapping rate (OR) as the basic metrics. CLE is the Euclidean distance between the center of the tracking result and the ground truth for each frame. OR is the overlap score between the two boxes, tracked bounding box $ROI_T$ and ground truth bounding box $ROI_{GT}$, which is computed based on the PASCAL challenge object detection score [58]:

$$OR = \frac{\text{area}(ROI_T \cap ROI_{GT})}{\text{area}(ROI_T \cup ROI_{GT})}$$

(24)

To rank the tracking performance, we compute the average CLE and OR across all frames of each image sequence.

1) **Comparisons Based on OR and CLE:** First, we compute the average CLE and OR of all image sequences on OTB50 and OTB100 and compare them with those of the other 15 state-of-the-art trackers, as shown in Table II. Generally, the proposed MCTTC achieves the best performance, with CLE of 26.61 pixels and OR of 61.17% on OTB50 and CLE of 36.26 pixels and OR of 55.37% on OTB100. Staple achieves the second-best performance on OTB50 both in terms of CLE (30.55 pixels) and OR (60.85%). For OTB100, KCF ranks second in terms of CLE (55.37 pixels), and DSST ranks second in terms of OR (52.61%).

### Table I

<table>
<thead>
<tr>
<th>Attribute</th>
<th>OTB50</th>
<th>OTB100</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>23</td>
<td>38</td>
</tr>
<tr>
<td>SV</td>
<td>28</td>
<td>63</td>
</tr>
<tr>
<td>OCC</td>
<td>29</td>
<td>48</td>
</tr>
<tr>
<td>DEF</td>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>MB</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>FM</td>
<td>17</td>
<td>41</td>
</tr>
<tr>
<td>IPR</td>
<td>31</td>
<td>49</td>
</tr>
<tr>
<td>OPR</td>
<td>39</td>
<td>62</td>
</tr>
<tr>
<td>OV</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>BC</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>LR</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>
Moreover, we perform quantitative comparisons with seven other state-of-the-art trackers on 18 challenging image sequences. The selected trackers are four CF trackers (KCF, DSST, STC, and CSK) and three part-based trackers (SCM, Struck, and IVT). The selected image sequences, mostly having occlusions or deformation. As shown in Tables III and IV, the proposed MCTTC outperforms the other trackers in terms of CLE and OR. The superiority of our method can be attributed to the effective incorporation of CFs and part-based representation in our framework.

Notably, we do not perform comparisons for Staple, RPT, TGPR, and CNT on OTB100 because we cannot find their tracking result files. In addition, for RAPCF [38], we cannot perform overall comparisons on OTB50 and OTB100 because the codes and tracking results on OTB50 or OTB100 have not been released.

2) Precision and Success Plots of OPE: The distance precision rate (DP) and overlap success rate (OS) are generally adopted to measure the overall performance of a tracker. A DP is the ratio of the frames in which CLE is smaller than a threshold. OS is the ratio of the frames in which the OR is larger than a threshold. The distance precision at a distance threshold of 20 pixels is employed as the representative DP score, and the overlap precision at an overlap threshold of 0.5 is employed as the representative OS score.

The DP and OS curves in one-pass-evaluate (OPE) [11], [25], namely, the precision and success plots of OPE, are used to present the overall performance of trackers.

Overall quality plots. Fig. 6 shows the overall quality plots of several state-of-the-art trackers on OTB50 and OTB100. With regard to representative DP and OS scores, the proposed MCTTC shows the best performance in the two datasets compared with the other trackers. The proposed method has a DP of 82.5% and an OS of 60.4% for OTB50, and a DP of 75.1% and an OS of 54.7% for OTB100.
Besides, we conduct experiments to evaluate how each component affects the tracker’s performance.

**Evaluations on component analysis.** The framework of MCTTC consists of three components, namely, part selection scheme, multi-part correlation tracker, and triangle-structure model. The proposed multi-part correlation tracker is the body and the basic tracker, denoted by MCT. Therefore, during the component analysis, we mainly conduct comparisons on part selection scheme and triangle-structure model, respectively.

How to select discriminative parts is important for part-based models. To evaluate the effectiveness of our part selection scheme, we adopt fixed divisions to determine local parts; therefore, the target is divided into multiple local parts with the same size and without overlaps. As shown in Fig. 8, we use five fixed division ways to generate the local parts for comparison. The corresponding evaluations are presented in the bottom row of Fig. 8. As shown in the figure, the proposed part selection scheme is the most effective among all of the fixed division ways to generate local parts.

The triangle-structure model is developed to describe the spatial relationship among parts, which is actually independent of the basic tracker. Therefore, the effectiveness of the triangle model can be verified by evaluating how the tracker performs when the structural similarity is introduced. As shown in Fig. 9, when the structural similarity obtained by our triangle model is ignored, MCTTC without triangle-structure constraints, that is, MCT, shows poor performance. Hence, the effectiveness of the proposed triangle-structure model is demonstrated.

**Attribute-based evaluation.** As indicated in Table I, OTB50 and OTB100 have 11 attributes. Therefore, we analyze tracking performance based on attribute evaluations for OTB50 and OTB100. Given the limited space of a page, we only present four attribute evaluations of the two datasets, namely, occlusions, out-of-plane, deformation, and scale variations, as shown in Figs. 10 and 11. Notably, the proposed MCTTC outperforms the other methods in the occlusions attribute in OTB50 and OTB100, which verifies that the problem of partial occlusion is addressed. In OTB50, the proposed MCTTC achieves at least the top two performance in most attributes. While in the scale variation attribute in OTB50, no specific tracker ranks first both in precision and success plots. The top four trackers are the proposed MCTTC, Staple, RPT, and DSST. The effectiveness of our MCTTC can be attributed to the use of the framework of particle filter for sampling the candidates, in which we apply six affine parameters, including a scale variable, to determine object motion. Another tracker, DSST, is robust and effective in handling scale variation because it was originally exclusively developed for this purpose.

In OTB100, the proposed MCTTC outperforms the other trackers in all attribute evaluations but the low resolution attribute. In both datasets, the evaluation results in the low resolution attribute show that the proposed tracker cannot only ranks fourth to second in precision and success plots. The reduced effectiveness and robustness at low resolution in the two datasets are due to the fact that we exploit holistic and local appearance to model the target. If the target is of low resolution, then the local part will be too small to contain sufficient information, resulting in the part filter not being robust enough during the tracking process.

Overall, the proposed MCTTC outperforms the other state-of-the-art trackers in the attribute-based evaluations, especially in the occlusions and out-of-plane rotation attributes.
C. Qualitative Evaluation

Fig. 12 shows the tracking results of the proposed MCTTC compared with those of the other top nine trackers (Staple, RPT, KCF, DSST, CNT, TGPR, Struck, SCM, and TLD) in our evaluations using several challenging sequences. Generally, these trackers perform well, and we conduct a qualitative analysis of them by using different attributes.

**Occlusions and deformation.** In the basketball sequence, in which the target undergoes occlusions and deformation, Struck, SCM, and TLD cannot track the player when he is occluded by another player. Only MCTTC, CNT, and TLD can capture the girl when she reoccurs from the occlusions of the pillar (jogging-1), due to the following reasons. The proposed MCTTC contains an adaptive update model scheme and part-based representation. CNT learns a robust feature descriptor, and TLD is a long-term tracker with a re-detection module. In the Surf sequence, although no deformation or other challenges, but the car undergoes a long-term occlusion by trees, and TGPR, TLD, and Struck cannot capture the target in whole frames. DSST, SCM, TLD, and Struck tend to drift away when heavy occlusions and deformation occur in david3.

**Scale variations, background clutter, and occlusions.** The woman sequence involves heavy partial occlusions and scale variations. Although most of the trackers can locate the target well in most frames but lose the target in several frames, all of them cannot adapt to the scale variations well. In the walking2 sequence, MCTTC, Staple, SCM, and CNT can lock the target after the woman is occluded by the man, and only the proposed MCTTC and SCM can capture the scale variations because...
Fig. 12. Tracking results of the top 10 trackers on 12 challenging sequences (from left to right and top to down are bolt, basketball, suv, jogging-1, david3, woman, walking2, liquor, tiger1, singer2, lemming, skating1). Frame indexes are shown in the top left of each figure. The results of the proposed MCTTC are viewed in red color.

both trackers use particle filters to model object motion. For the liquor sequence involving scale variations, occlusions, and background clutter (i.e., the background near the target has a similar color or texture as the target), MCTTC, Staple, RPT and KCF perform better than the others in terms of locating the target. In adapting to the scale variation, the proposed MCTTC shows the best performance.

Illumination changes, scale variations, and fast motion. The sequences tiger1, singer2, lemming, and skating1, involve illumination changes, and the proposed MCTTC performs well in all these challenging sequences. However, the other trackers possess several problems as follows. In tiger1 wherein the target undergoes scale variations, TLD, CNT, Struck, SCM, TGPR, and DSST cannot keep locating the object all over the frames. In addition, Struck, TLD, and SCM show reduced effectiveness in the singer2 sequence that involves the same challenges as tiger1. Except for MCTTC, only TLD and Struck can track the target to the end in lemming, in which the target simultaneously undergoes scale variations, out-of-plane rotation, and fast motion. Even when the skating1 sequence involves illumination changes, scale variations, deformation, fast motion and out-of-plane rotation, MCTTC, Staple, DSST, RPT, KCF, and CNT can perform well.

The qualitative evaluation generally demonstrates that the proposed MCTTC is more competitive than the nine other state-of-the-art trackers.

VI. Conclusions

In this paper, we propose a novel multi-part correlation tracker with triangle-structure constraints. By training multiple correlation filters (CFs) for one holistic object and multiple local parts, these CFs are employed to estimate the target by jointly measuring the correlations of the holistic candidate and its local parts. Due to the divide-and-rule strategy, the model can effectively handle partial occlusions. Meanwhile, preserving the holistic presentation contributes remarkably when local filters drift away in a specific scene with low resolution. Considering that the triangle is the most stable structure among all the graphs, we employ several triangles to describe the spatial relationship among parts to model the structure of the target. During the tracking process, the structural similarity of a candidate is estimated by measuring how similar the estimated triangles formed in the candidate are to the ones formed in the original target. Moreover, an effective part selection scheme is introduced based on energy and integrity, which is generally applicable to part-based tracking models. Extensive experiments on benchmarks demonstrate our method outperforms the state-of-the-art methods.

VII. Acknowledgement

This work is supported by National Nature Science Foundation of China (No. U1611461, 61876135, 61876086), National Key R&D Program of China (No. 2017YFC0803700), Hubei Province Technological Innovation Major Project (2017AAA123, 2018AAA062), and Nature Science Foundation of Jiangsu Province (No. BK20160386).
REFERENCES


Weijian Ruan received the B.E. degree from Electronic Information School of Wuhan University in 2014. He is currently pursuing his Ph.D. degree in school of computer science, Wuhan University. From 2018.03 to 2018.09, he had been an intern in the National Institute of Informatics, Tokyo, Japan. His research interests focus on computer vision, multimedia analysis and machine learning, where he has published several conference papers including ICME, AAAI, ICIP, and ICNN.

Jun Chen received the M.S. degree in Instrumentation from Huazhong University of Science and Technology, Wuhan, China, in 1997, and Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2008. Dr. Chen is the deputy director of National Engineering Research Center for Multimedia Software, and a professor in school of computer science, Wuhan University. His research interests include multimedia analysis, computer vision and security emergency information processing, where he has published over 50 papers.

Yi Wu received the Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. Then he joined Nanjing University of Information Science and Technology, Nanjing, China. In 2016, he joined Nanjing Audit University as the Runze Professor. He did his postdoctoral training at Temple University, USA, University of California, Merced, USA and The Ohio State University, USA. He was a research faculty in Indiana University School of Medicine. His research interests include computer vision, medical image analysis, and machine learning. His tracking benchmark is one of the most popular tracking evaluation platforms.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently a Professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, object detection and tracking, and intelligent video surveillance.

Chao Liang received the Ph.D degree from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012. He is currently working as an associate professor at National Engineering Research Center for Multimedia Software (NERCM-S), Computer School of Wuhan University, Wuhan, China. His research interests focus on multimedia content analysis and retrieval, computer vision and pattern recognition, where he has published over 60 papers, including premier conferences such as CVPR, ACM MM, AAAI, IJCAI and honorable journals like TNNLS, TMM and TCSVT, and won the best paper award of PCM 2014.

Ruimin Hu received the B.S. and M.S. degrees from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and in 1990, and Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994. Dr. Hu is a professor of School of Computer Science, Wuhan University. He has published two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance and multimedia processing and analysis.

Junjun Jiang received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014. From 2015 to 2018, he was an Associate Professor at China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. He won the Finalist of the World’s FIRST 10K Best Paper Award at ICME 2017, and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award. His research interests include image processing and computer vision.