Research Article

Optimal Metric Evaluation-Based Multicue Inverse Sparse Appearance Model for Object Tracking

Xiaowei An ¹, Qi Zhao ², Nongliang Sun ³ and Quanquan Liang ³

¹College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China

²School of Aviation Engineering, Shandong Polytechnic College, Jining 272067, China

³College of Electronics and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Correspondence should be addressed to Quanquan Liang; quanquan406@gmail.com

Received 1 July 2020; Revised 29 October 2020; Accepted 28 November 2020; Published 16 December 2020

Academic Editor: Stylianos Georgantzinos

Copyright © 2020 Xiaowei An et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to obtain the discriminative compact appearance model for tracking objects effectively, this paper proposes a new structural tracking strategy that includes multicue inverse sparse appearance model and optimal metric evaluation between online robust templates and a limited number of particle samples in the looping process. Multicue inverse sparse appearance model globally improves the efficient selection of informative particle samples that can avoid the cumbersome coding and decoding cost for the trivial random particle samples. Only the most potential crucial cases are involved in each tracking loop. This refrains from unreasonable, rough numerical reduction of particle samples and also keeps the unbiasedness and dynamic stochasticness of the sampling process. Meanwhile, low-rank self-representatives for positive and negative samples facilitate the formulation of a suitable code book that arranges the useful sparse coefficients for feature bags and facilitates optimal metric evaluation for online training. It also alleviates the accuracy degradation of tracking occluded objects and improves the robustness of the tracker. Both of them preserve the discriminative compactness of target which speeds up particle filtering localization to separate the target object from distractors. Moreover, the proposed method exploits online appearance representations to learn the sharing compact information that avoids massive calculation burdens for massive visual data.

1. Introduction

As an effective solution to locate the interesting target, object tracking is seamlessly deployed in several surveillance services, which is very necessary to acquire better optimal appearance modeling method to satisfy the distributed surveillance requirements, such as the compactness of the model and low computation cost of transmission. Despite the number of solutions having been implemented in this field, it also often accompanies with challenging problems about the object appearance model, such as occlusions, illumination changing, and pose variations [1]. For the reason that high-accuracy surveillance needs expensive and complex deployment under limited computation resources in the real environment, it is crucial to leverage compactness and robustness for appearance modeling. Through machine learning methods for robust improvements, incremental

subspace learning was utilized to tackle with the templates' dramatic changes [2–4] in order to alleviate dirty templates' training. They are not only, to some extent, effective to exploit the intrinsic subspace structure but also could not avoid huge storage of high-dimension data for nuclear norm minimization. To fix up this drawback, algorithms [5, 6] with sparsity representation were presented by the multilinear framework under the minimization of reconstruction error. However, the learned sparse appearance model could not provide enough spatial context information for the reason that sparsity representation coefficients were often arranged for target samples in each tracking loop individually. It was easy to ignore global constraints on the related subspace structures among the whole video sequences. Meanwhile, sparse decompositions in accumulated looping also consume higher time that results in low running rate and high energy requirement. Cooperative sparse appearance model

that owned the sparse generative model (SGM) and sparse discriminative classifier (SDC) utilized the global templates to update the appearance model and measure the similarity through trivial discriminative blocks [7]. Methods [8–10] took partial and spatial information into consideration to exploit more robust templates with massive burdens about atom dictionary construction and pooling calculation procedures. Even wavelet transformation-based features were extracted to improve the reliability of appearance modeling where joint dictionaries for sparse coding still required tough storage procedures [11]. Intuitively, discriminative method provides adaptive complementary option for various appearance changes [1]. Good discriminative representations usually need mounts of supervised labels to fit the real data distribution [12–14]. Different classifiers were utilized to obtain more discriminative appearance models. Method in [15] relied on the background information where the most discriminating metrics for tracking were classified to keep more stable tracking. In [16], the random forest-based online multiview semisupervised learning algorithm that updated subtrees with individual labels for the unlabeled data was provided. The Hough forest-based back-projection was used in [17] to generate the structural patches. The spatial regularization was adopted in [18] to penalize the learning classifier. Bootstrapped sequential states between frames were shown in [19] to avoid random samples contaminating labeled examples. Recently, convolutional neural network-based methods captured ample information that was sophisticated in describing appearance models by the multilayer nonlinear transformations [20–23]. However, pooling procedure-based abstract convolutional features from network layers might ignore the original complex feature attributes of the appearance model inside the image structure. Furthermore, even utilization of transfer learning was able to adopt large sets of pretraining data, but complex high cost was also spent in the visual data collection, annotation labelling, and training ground-truth data. In our perspective, it was challenging to pursue for deep appearance models under the limited samples. Also, the transfer learning model which resulted from the large-scale dataset might have a certain divergence among various domains [22, 24].

To explore the robust and discriminative compact appearance model and to alleviate the heavy calculation cost in the looping process, this paper proposes an optimal structural tracking strategy that consists of global sparse representatives and local sparse coding feature bag-based optimal metric evaluation. In summary, the main contributions include the following:

- (1) Multicue inverse sparse appearance model avoids the heavy computation of the redundant particle sampling which results from the trivial random procedures. It obtains the most informative particle cases for structural appearance modeling.
- (2) Positive and negative samples are replaced by suitable sparse coefficient-based feature bags in the local level that can yield the optimal metric composition. It is potentially better suited for matching evaluation

by limited potential powerful information in the spare coding phase. Also, this way reserves the target subtly discriminativeness of the compact model.

2. Background Information

2.1. Inverse Sparse Appearance Model. Given normalized M candidates $\{X_{t+1}^M\}$ by particle filtering in the $(t + 1)$ -th frame, previous target region ROI_t as the template $\{X_t^T\} \in R^{d \times 1}$ in the t -th frame can be coded by the dictionary [25, 26]: $\{D_{t+1}\} = [X_{t+1}^1, X_{t+1}^2, \dots, X_{t+1}^M] \in R^{d \times M}$. Afterwards, sparse decomposition of template ROI_t is presented by nonnegative combination of sparse coefficients $\{b_{t+1}^*\} = [b_{t+1}^1, b_{t+1}^2, \dots, b_{t+1}^M] \in R^{1 \times M}$, while template reconstruction error achieves the minimum constraint with penalty term λ as shown in the following equation:

$$\begin{aligned} \arg \min & \|X_t^T - D_{t+1} b_{t+1}^{*T}\| + \lambda \|b_{t+1}^*\|_1, \\ \text{s.t. } & b_{t+1}^* \geq 0. \end{aligned} \quad (1)$$

2.2. Low-Rank Self-Representatives. Optimal selection of low-rank exemplar representatives for high-dimensional data structure is efficiently described by the relevant data groups [27]. Attributes of self-representation in such cases of high relevance are exploited in order to obtain the most crucial ones. Given N data samples in a dataset $\mathbb{R}^m$ as columns of data matrix $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{M \times N}$, the optimization problem is shown as follows:

$$\begin{aligned} C &= \arg \min \|X - XC\|_F^2, \\ \text{s.t. } & \|C\|_{0,q} \leq n. \end{aligned} \quad (2)$$

Here, $C = [c^1, \dots, c^N] \in \mathbb{R}^{N \times N}$ is the coefficient matrix and $\|C\|_{0,q}$ counts n ($\ll N$) nonzero rows of C . This compact learning process can be treated as a self-representative procedure that is the analogous structural representation of original data.

2.3. Metric Evaluation. Metric evaluation is calculated by the difference optimization between two feature vectors that can be defined by the positive semidefinite matrix. Mahalanobis distance [28] between $\mathbf{x}_i$ and $\mathbf{x}_j$ is a very famous metric evaluation as shown in the following equation:

$$d_M^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j). \quad (3)$$

Here, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in R^{d \times n}$ represents n vectors $\{\mathbf{x}_i\} \in R^{d \times 1}$ of the training sets. Afterwards, M matrix is factorized into the positive semidefinite matrix $M = A^T A$. So, equation (3) is transformed into a new style as follows:

$$\begin{aligned} d_M^2(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} \mathbf{A}^T (\mathbf{x}_i - \mathbf{x}_j), \\ &= \|\mathbf{A}^T \mathbf{x}_i - \mathbf{A}^T \mathbf{x}_j\|_2^2. \end{aligned} \quad (4)$$

Online learning matrix $M(A)$ facilitates the mapping transformation of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ into a new low-dimensional

subspace. This way also takes a new feasible distance metric instead of the original one.

3. Proposed Algorithm

3.1. Multicue Inverse Sparse Appearance Model-Based Particle Sampling (MISAMPS). Random sampling procedures preserve the stochastic evaluation attributes of nonlinearity and analysis uncertainty, but for large mounts of random samples, computation cost is still a big problem under the limited resources. In order to alleviate the redundancy that results from random particle sampling, this paper applies the global-level multicue inverse sparse appearance model to select the most powerful particle exemplars in the tracking loop. After the initial extraction of ROI (region of interest) in the t -th frame, the normalized N random particle sampling states $\{X_{t+1}^i\}_{i=1,2,3,\dots,N}$ in the $(t+1)$ -th frame are firstly segmented into $\tilde{K}$ local patches that can be coded for the atom dictionary. Besides, the segmented patches are arranged for various weights $\{w^i\}_{i=1,2,3,\dots,\tilde{K}}$ that exploit the potential connectivity for separating the foreground object from the background discriminatively when the ROI faces partial occlusion situations. In this paper, multicues (Gray/Hog/Lbp) of each patch, respectively, describe the inverse sparse appearance model for robust representations. As shown in Figure 1, compact selection of powerful particle samples by inverse sparse representation copes with the original sampling of the redundant structure; therefore, it is more valuable to consider the limited sampling cases that are obtained after multicue inverse sparse modeling. To alleviate the tracking drifting problem, the uniform corresponding patches are processed by each single cue extraction, respectively, that can pursue for more accurate local weight distribution.

Commonly, local-level patch weight distribution can be gradually optimized by the adaptive AdaBoost process [25] or quadratic programming theory [29]. For the adaptive weight distribution, the size of each normalized patch is set as $WW \times HH$ (pixel). Coordinate (u_i, v_i) exists inside the i -th patch ($i \in 1, \dots, \tilde{K}$). Following the previous $\{w_{\text{type}}^i\}_{\text{type}=\text{Gray/Hog/Lbp}}$ distribution in the previous t -th frame, the weight distribution for partial occlusion is shown in Figure 2. This exhaustive presentation ensures that the evolution of feature confidence consistently exists between the current frame and the previous frame. It seems obvious that multicue-based structural weight distributions reflect the dynamical confident patches' arrangement in which the no-occluded patches (warm-color patches) show high weight distribution and vice versa. Considering the potential structural diversity among different features in the target ROI, multicues provide more various optimal particles' proposals with low formidable procedures. If normalized M candidates $\{X_{t+1}^{1,\dots,M}\}$ are sampled in the $(t+1)$ -th frame, storage of previous t target ROIs as the templates $X_t^{(j/\text{type})} \in R^{d \times M}$ $j=1,\dots,t|\text{type} \in \{\text{Gray}, \text{Hog or Lbp}\}$ until the t -th frame can be coded by the respective dictionary: $\{D_{t+1}^{\text{type}}\} =$

$\{X_{t+1}^{(1/\text{type})}, X_{t+1}^{(2/\text{type})}, \dots, X_{t+1}^{(M/\text{type})}\} \in R^{d \times M}$. According to equation (1), sparse decomposition for the template by nonnegative sparse coefficients $\{b_{t+1}^{\text{type}}\}$ is implemented with

the help of $\{w_{\text{type}}^i\}_{i \in \tilde{K}\text{-th patch}}^{\text{type}=\text{Gray/Hog/Lbp}}$ distribution until template reconstruction error achieves the minimum constraint under the penalty term λ^{type} constraint as shown in the following equation:

$$\begin{aligned} \arg \min & \|w_{\text{type}}^i \odot \{X_t^{(j/\text{type})} - D_{t+1}^{\text{type}} b_{t+1}^{\text{type}}\}\| + \lambda^{\text{type}} \|b_{t+1}^{\text{type}}\|_1, \\ \text{s.t. } & \{b_{t+1}^{\text{type}}\} = [b_{t+1}^{(1/\text{type})}, b_{t+1}^{(2/\text{type})}, \dots, b_{t+1}^{(M/\text{type})}] \in R^{1 \times M}, \\ & b_{t+1}^{\text{style}} \geq 0. \end{aligned} \quad (5)$$

3.2. Feature Bag-Based Optimal Metric Evaluation (FBOME). To select the most suitable result among the provided sets of particle exemplars, the feature bags are processed in advance which own more discriminating attributes than the original color feature-based representation. Instead of coding by multiple patch-based K -means clustering for the convolutional filtering bank in the tracking loop, the principle atom-based code book (PACB) employs the low-rank self-representatives to represent the imperative atoms as shown in Figure 3 that can bring about the subsequent feasible sparse labels for the whole ROI area. Given N templates $\{T_{t+1}^i\}_{i=1,2,3,\dots,\tilde{K}}$ in the $(t+1)$ -th frame, column vectorization sets $\{\mathbf{t}_{t+1}^i\} \in R^{\text{sizB} \times N}$ can be decomposed under the minimal reconstruction error with (l_0/l_q) constraints. sizB equates $WW \times HH \times \tilde{K}$. According to equation (2), $n (\ll N)$ representatives $\{\mathbf{t}_{t+1}^i\} \in R^{\text{sizB} \times n}$ are selected as principle atoms to take instead of original $\{\mathbf{t}_{t+1}^i\}_{i=1,2,3,\dots,\tilde{K}}$ cases. $\|C\|_{0,q}$ in equation (2) can be shown as

$$\|C\|_{0,q} = \sum_{i=1}^N I(\|c^i\|_q > 0), \quad (6)$$

where c^i is the i -th row of C and $I(\cdot)$ is the indicator that shows the number of nonzero entry rows of C . Its corresponding i -th columns are the nonpowerful representatives for the whole $\{\mathbf{t}_{t+1}^i\} \in R^{n \times \text{sizB}}$ structure [27]. To solve the NP-hard problem [30] for the l_0 -norm constraint problem, l_1 -norm is usually applied for the new limitation ($1^T C = 1^T$) concerning the elements $\{c_{ts}\}$ of C . So, (2) can be solved by the following equation:

$$\begin{aligned} C &= \arg \min \|\{\mathbf{t}_{t+1}^i\} - \{\mathbf{t}_{t+1}^i\} \cdot C\|_F^2, \\ \text{s.t. } & \|C\|_{1,q} \leq \sigma, 1^T C = 1^T. \end{aligned} \quad (7)$$

Here, σ is a nonnegative parameter, and $q > 1$ confirms the convex optimization. $\|C\|_{1,q} = \sum_{j=1}^N \|c^j\|_q$ summarizes the l_q -norms of rows of C . This solution describes the crucial sets of representatives for the related rows in the data structure. More nonzero entries in the i -th rows of C play higher imperative weights in the data self-representation. We can

FIGURE 1: Illustration of the proposed algorithm.

FIGURE 2: Adaptive weight distribution for local patches.

FIGURE 3: Imperative atom representation by low-rank self-representatives.

obtain $\|c^1\|_q < \|c^2\|_q < \|c^3\|_q < \dots < \|c^n\|_q$. Such solutions which are mentioned above compress the redundant information of $\{T_{t+1}^i\} \in R^{K \times N}$ with n low-rank PACB sets $(D_{PACB}) \{\tilde{T}_{t+1}^i\} \in R^{K \times n}$ efficiently.

Afterwards, nonzero $\{b_{t+1}^{*type}\} \in R^{1 \times s}$ ($s < M$) corresponding powerful cases $\{X_{t+1}^{(1/type)}, \dots, X_{t+1}^{(s/type)}\} \in R^{K \times s}$ from the multicue inverse sparse model are coded to preserve more vital, potential, structural information by the obtained dictionary D_{PACB} that can describe the spatial appearance variations appropriately for the simultaneous tracking process. If column factorization of j -th case $X_{t+1}^{(j/type)}$ is named as $\mathbf{x}_{t+1}^{(j/type)} \in R^{sizB}$, $\beta_j \in R^{1 \times n}$ are the bag of words (BOW) [31] which sparsely code the mapping process in equation (8). Here, the least absolute shrinkage and selection operator (LASSO) solution [32] is implemented with the greedy looping procedures in that the algorithm updates atoms β_j until the residual is lower than the initialized threshold or until enough atoms are obtained.

$$\begin{aligned} \min_{\beta_j} & \|\mathbf{x}_{t+1}^{(j/type)} - D_{PACB}\beta_j\|_2 + \lambda|\beta_j|, \\ \text{s.t. } & \beta_j \geq 0. \end{aligned} \quad (8)$$

As shown in Figure 4, structural pyramid-pooling procedures, respectively, sample sparse coefficient-based feature bags (BOW) covering the whole ROI with different sampling sizes of $\epsilon^2_{|\epsilon \in \{1,2,3\}}|$, and then average pooling for sparse coefficients in each subset is calculated according to the ϵ -th level. Finally, pooling results in every level are concatenated in the linear pooling way. This manner preserves the corresponding sparse statistics of spatial intensity among various scales inside the pyramid structure. Meanwhile, online training templates include both positive sampling cases and consecutive recognition results which guarantee the consistent property of the target ROI in which spatial information of temporal templates can be encoded suitably. Therefore, the optimal metric evaluation is triggered by sparse coefficient-based feature bags from dictionary PACB which ensures the online dynamic updating templates during the whole tracking interaction. Given the target template $\mathbf{T}_t \in R^k$ of the t -th frame and the q -th sampling candidate $\{\bar{\mathbf{C}}_{t+1}^q\} \in R^k$ in the $(t+1)$ frame, intuitively, Mahalanobis distance between them can be described as equation (9) according to equations (3) and (4):

$$\begin{aligned} d_M^2(\bar{\mathbf{C}}_{t+1}^q, \mathbf{T}_t) &= (\bar{\mathbf{C}}_{t+1}^q - \mathbf{T}_t)^T \mathbf{M}(\bar{\mathbf{C}}_{t+1}^q - \mathbf{T}_t), \\ &= (\bar{\mathbf{C}}_{t+1}^q - \mathbf{T}_t)^T \mathbf{A} \mathbf{A}^T (\bar{\mathbf{C}}_{t+1}^q - \mathbf{T}_t) \\ &= \|\mathbf{A}^T \bar{\mathbf{C}}_{t+1}^q - \mathbf{A}^T \mathbf{T}_t\|_2^2. \end{aligned} \quad (9)$$

Here, $\mathbf{A} \in R^{k \times k}$ is the symmetric positive semidefinite matrix that can be iteratively determined by the online metric learning method [33] during the tracking process. However, with more random sampling candidates in each tracking looping procedure, computation cost spent is obviously much higher. Moreover, the variations between redundant candidates may affect the adjustment for metric

evaluations among the following pairs of samples. In order to improve metric evaluation robustly, this paper takes the sparse coefficient-based feature bags instead of original features.

$$\begin{aligned} d_M^2(\bar{\mathbf{C}}_{t+1}^q, \bar{\mathbf{B}}_{Tt}) &= (\bar{\mathbf{C}}_{t+1}^q - \bar{\mathbf{B}}_{Tt})^T \beta_{A_{t+1}} \beta_{A_{t+1}}^T (\bar{\mathbf{C}}_{t+1}^q - \bar{\mathbf{B}}_{Tt}), \\ &= \|\beta_{A_{t+1}}^T \bar{\mathbf{C}}_{t+1}^q - \beta_{A_{t+1}}^T \bar{\mathbf{B}}_{Tt}\|_2^2. \end{aligned} \quad (10)$$

Here, $\bar{\mathbf{C}}_{t+1}^q$ and $\bar{\mathbf{B}}_{Tt}$ are the pyramid-pooling structure-based sparse coefficient sets for the q -th candidate $\bar{\mathbf{C}}_{t+1}^q$ and template $\mathbf{T}_t$, respectively. Meanwhile, $\{\beta_{A_{t+1}}\}$ (or $\{\beta_{A_{t+1}}\}$) results from temporal multi-instance metric learning with positive sets $\{\mathbf{P}_t^s\}$ and negative cases $\{\mathbf{N}_t^s\}$ by automatic shift sampling selection around the previous target ROI in the t -th frame. Furthermore, part-based low-rank self-representation decomposition is employed here again to extract more informative positive $\{\mathbf{P}_t^{sl}\}$ and negative $\{\mathbf{N}_t^{sl}\}$ selections for more robust training. Therefore, the constraints are limited as follows: the difference in the t -th frame between positive cases $\mathbf{P}_t^{sl}$ and the template $\mathbf{T}_t$ is more than or equal to a small value ς .

$$|\mathbf{P}_t^{sl} - \mathbf{T}_t| \geq \varsigma. \quad (11)$$

The distance between consecutive elements of template $\{\bar{\mathbf{B}}_{Tt}\}$ should be less than a small value ς .

$$|\bar{\mathbf{B}}_{T(t+1)} - \bar{\mathbf{B}}_{Tt}| < \varsigma. \quad (12)$$

The difference in the t -th frame between positive cases $\mathbf{P}_t^{sl}$ and negative cases $\mathbf{N}_t^{sl}$ should be a large margin.

$$|\mathbf{P}_t^{sl} - \mathbf{N}_t^{sl}| \gg \varsigma. \quad (13)$$

$\beta_{A_{t+1}}$ is solved by temporal metric learning with the pair label $\bar{L}$ and the LogDet optimization [33] under constraint parameter $\bar{\lambda}$. Thus, it can be seemed as the following equation:

$$\begin{aligned} \beta_{A_{t+1}} &= \beta_{A_t} + \bar{\lambda} \beta_{A_t} (\beta_{u_i} - \beta_{v_j}) (\beta_{u_i} - \beta_{v_j})^T \beta_{A_t}, \\ \{\beta_{u_i}, \beta_{v_j}, L\} &\in \{\mathbf{T}_t, \mathbf{P}_t^{sl}, +1\} \text{ or } \{\mathbf{P}_t^{sl}, \mathbf{N}_t^{sl}, -1\}. \end{aligned} \quad (14)$$

Similarly, $\beta_{\mathbf{P}_t^{sl}}$ and $\beta_{\mathbf{N}_t^{sl}}$ are also defined as the pyramid-pooling structure-based sparse codes for $\mathbf{P}_t^{sl}$ and $\mathbf{N}_t^{sl}$, respectively. With the above optimal metric representation, object tracking process can be treated as selecting the most similar candidate from the limited sampling sets that have been provided by the multicue inverse sparse appearance model. Under the Bayesian inference framework [34], the likelihood $p(\mathbf{Y}_{t+1} | \mathbf{s}_{t+1})$ can be defined as follows:

$$p(\mathbf{Y}_{t+1} | \mathbf{s}_{t+1}) \propto -\exp(d_M^2(\bar{\mathbf{C}}_{t+1}^q, \bar{\mathbf{B}}_{Tt})). \quad (15)$$

3.3. Model Update Mechanism. Within the iterative process for updating the online template library, the proposed algorithm generates both positive and negative samples in a

FIGURE 4: Feature bag representation and pyramid mapping structure.

certain radius that can be treated as the auxiliary samples to jointly train the model for antidistractions from background trivial cluttering. All the positive samples are located inside the radius r_1 which is near the positive label instance; likewise, negative ones stay in the interval $r_1 < r_2$ at a certain distance far from positive label instances. Meanwhile, the proposed algorithm does not only update the template $\{T_{t+1}\}$ through pyramid-pooling structure procedures (11)~(14) but also adaptively refreshes the templates based on reconstruction error Err with the threshold thr and an adaptive tuning parameter μ :

$$\begin{aligned} \text{Err} &= \|T_t - (w_{\text{type}}^i \odot D_{t+1}^{\text{type}} b_{t+1}^{\text{type}})\|_2, \\ T_{t+1} &= (1 - \mu)T_t + \mu(w_{\text{type}}^i \odot D_{t+1}^{\text{type}} b_{t+1}^{\text{type}}), \\ \text{s.t. } \text{Err} &> \text{thr}. \end{aligned} \quad (16)$$

4. Experiment and Analysis

The experiments are carried on a PC with Intel i7-2.60 GHz CPU and 8 GB storage with MATLAB implementation. All the parameter settings in the experiment are normalized for fair comparisons. L_1 minimization optimization is solved by SPAMS package [35] with the regularization constant $\lambda^{\text{type}} = 0.13$. For a good tradeoff between effectiveness and time cost, two hundred particles are randomly sampled for providing enough candidates in each tracking loop. All target ROI areas are initialized by manual and modelled as previous sections. Local patches are normalized for 32×32 pixel size for the affine transformation. The relation of intervals r_1 and r_2 is commonly $r_2 = 2\sqrt{2} \times r_1$. The model update rate μ and error threshold thr in equation (16) are 0.95 and 0.065, respectively. To demonstrate the robustness of the proposed algorithm, we firstly give several basic experimental comparisons involving different metric evaluation methods related to the proposed algorithm. Secondly, we adopt the classic OTB database [1] to give the comparisons between our tracker and others.

4.1. Basic Metric Evaluation. OPF [34]: original particle filtering without sparse coding modeling; CBIS [36]:

convolutional block feature-based metric evaluation with the inverse sparse appearance model; ASLA [8]: original block feature-based metric evaluation with the sparse appearance model; OBIS [25]: original block feature-based metric evaluation with the inverse sparse appearance model. Test video frames involve several real-life target-tracking tasks, such as low-contrast environment, sharp lighting influence, shape changing, scale variation, cluttering background, and fast pose movement.

Figure 5(a) presents the tracking results under the low-resolution condition. It shows that the human target is tracked successfully at the early stage by almost all the algorithms. However, OPF, ASLA, and OBIS trackers fail when the other pedestrians walk cross the target's directional routine. It also proves that only feature-based metric comparison lacks enough stable attributes than the hierarchical feature structures, such as the convolutional features or our sparse coefficient-based pyramid-pooling structure. Meanwhile, even the sparse modeling procedure-based OBIS tracker also may lose the target as shown in Figures 6(a) and 7(a). Linear combination-based final sample selection will accumulate the tracking discrepancy among sets of looping calculations which degenerate the tracking accuracy. Figure 5(b) presents that various illuminations with blurring in the ROI produce more troubles for the tracking process. As shown in Figures 6(b) and 7(b), the CBIS tracker and proposed algorithm play more better roles before the 42-nd frame that the biker ROI faces severe lighting. Then, CBIS tracker drifts the correct target area. For the background cluttering situation shown in Figure 5(c), it is trivial to track sharp activity in this squashed environment. Although ASLA and CBIS trackers almost capture a certain part of the target, their effective overlapping area is still less than the proposed algorithm. For the shape variation case shown in Figure 5(d), waving T-shirt is not easy to provide the stable appearance to be followed. In spite of each tracker with the same adaptive scale parameters, other trackers are not robust enough to describe this appearance. Scale variations shown in Figure 5(e) are more obvious that camera view focuses on the singer face from near to far within the dark and lighting stages. Figures 6(e) and 7(e) show the center error representations and overlapping rate comparisons with other trackers. CBIS, OBIS, and ASLA perform steadily before the 80-th frame under the dramatic lighting condition.

However, they all move within several different degrees when the singer face changes the directions. With the target scale declining, the proposed algorithm with double dynamic update schemes and optimal selection avoids the transient loss problems. For the sequences mottle face as shown in Figure 5(f), sunshine projects more mottle poles above the girl face with a certain plane of rotation. Even if there is no tracker to obtain very high overlapping rate through the whole sequence, the proposed tracker performs well in most of the stages. Differences in Figures 6(f) and 7(f) illustrate the accurate trend of all the trackers. In order to analyze the quantitative stability, the average center errors and the average overlapping rates of the five trackers in these experimental comparisons are calculated and shown in

— OPF
— OBIS
— CBIS

— ASLA
— Proposed algorithm

(a)

— OPF
— OBIS
— CBIS

— ASLA
— Proposed algorithm

(b)

— OPF
— OBIS
— CBIS

— ASLA
— Proposed algorithm

(c)

— OPF
— OBIS
— CBIS

— ASLA
— Proposed algorithm

(d)

— OPF
— OBIS
— CBIS

— ASLA
— Proposed algorithm

(e)

FIGURE 5: Continued.

FIGURE 5: Center error representations.

FIGURE 6: Center error representations.

Tables 1 and 2. They clearly display that the proposed algorithm has a good tracking performance in the vast majority of situations.

According to the comparison results, the proposed tracker ranks top two among all trackers in terms of optimal metric-based particle sampling selection. The CBIS tracker achieves the tracking task with the second lower average errors. It confirms the collaboration of the inverse sparse appearance model and discriminative particle sampling selection in the frequency domain through the K -means clustering-based dual-layer convolutional networks.

However, this method needs initial cluster number K for the construction of filter banks which may result in opaque target localization in the complex environments. Also, it does not consider the redundant sample computation cost in the dynamic update procedures; thereby, its further tracking performance is limited in a certain sense. For the ASLA tracker, it must calculate multiple sparse decompositions for each patch in each candidate. Then, each independent patch group must be evaluated through the max-pooling scheme. Thus, much computation memories are spent for coding the inefficient particle samples that hinder its real-time

FIGURE 7: Overlapping rate representations.

TABLE 1: Average center error.

Seqs	OPF	OBIS	CBIS	ASLA	Proposed
Pedestrian	56.2020	57.2070	6.9162	67.0526	7.0256
Bike	114.7249	49.8656	34.2210	65.6494	11.9540
Warriors	27.7086	23.1143	10.8556	13.5593	9.4310
T-shirt	49.1336	48.5074	25.4733	15.1385	13.7467
Singer	47.3966	22.7587	10.8144	12.7057	9.0016
Mottle	18.7783	18.4401	20.4053	17.6786	9.0359

TABLE 2: Average overlapping rate.

Seqs	OPF	OBIS	CBIS	ASLA	Proposed
Pedestrian	0.2153	0.1842	0.6487	0.1254	0.6130
Bike	0.0581	0.1170	0.2173	0.0352	0.5045
Warriors	0.3259	0.4253	0.5929	0.5904	0.7035
T-shirt	0.0526	0.2093	0.1793	0.4055	0.6274
Singer	0.1016	0.4820	0.6048	0.3696	0.6935
Mottle	0.4670	0.2901	0.4311	0.6002	0.6501

application. Although the OBIS and OPF trackers have much simpler structures and faster tracking speed than the proposed method, they do not own efficient and robust appearance models for test video sequences. Especially, OBIS adopts direct linear combination of particle samples rather than optimal selection. It is easy to drift the target for the reason that the accumulation of weak discrepancies will lead to degenerative states in consecutive procedures.

4.2. OTB Dataset Comparison. We also evaluate the proposed algorithm with other state-of-the-art algorithms including SRDCF [37], SAMF [38], HDT [39], DSST [40], KCF [41], SCM [7], L1APG [42], MIL [19], and CT [43] on the widely utilized OTB dataset. The benchmark dataset contains 50 different sequences with ground-truth annotation.

Unified categories of 11 challenging attributes are proposed here. Precision measures the center location error, which means the average difference between the center locations of targets and the ground truths. The final average center location error over all the frames of one sequence defines the overall performance. Here, we set the precision score for each tracker as the threshold of 20 pixels. The overlap ratio defines the overlapping relation between the predicted target area A_p and the ground-truth area A_g : $S = ((A_p \cap A_g) / (A_p \cup A_g))$; the final performance of a tracker on a sequence depends on the storage of successful frames in which S is more than a useful threshold. Under the success rate value at 0.5 threshold, the evaluation can be ranked by the area under the curve (AUC) of each success plot. The proposed comparison adopts the one-pass evaluation (OPE) throughout the whole sequence with the setting initialized by the first

FIGURE 8: (a) Precision plots and (b) success plots of OPE.

FIGURE 9: Continued.

FIGURE 9: Details of precision plots.

FIGURE 10: Details of success plots.

TABLE 3: Running speed (frames/second).

	SRDCF	SAMF	HDT	DSST	KCF	SCM	L1APG	MIL	CT	Ours
FPS	25	23	18	28	35	18	23	20	18	26

frame's ground-truth position. All the precision and success plots of the proposed comparison are shown in Figures 8(a) and 8(b). Different efficiencies of all trackers in various environmental attributes are illustrated in Figures 9 and 10. The proposed tracker (0.820/0.705) is in the top-2 precision plots and in the top-3 success plots, which obtained great improvements on the sparse-related trackers SCM, MIL, and L1APG. Also, it is better than some of correlation filter-based trackers, such as CT, KCF, DSST, and HDT. In order to fairly compare with the basic power of the tracker's structure, we only adopt the hand-craft features in all the trackers. From Figure 8(a), we can see that our method achieves at least top-2 highest precision rate in all challenging attributes except deformation and occlusion. In terms of deformation, the precision rate of the proposed tracker (73.0%) is just inferior to SRDCF (84%), KCF (80.4%), and HDT (79.0%). As to the occlusion, the precision rate achieved by our method (81.5%) is almost the same compared to the best score achieved by SRDCF (83.3%) and SAMF (82.8%) for the reason that SRDCF owns the more powerful discriminative kernel strategy which occupies much better positions than our proposed algorithm, but from Figure 10, our method achieves the highest precision rate in several challenging attributes such as low resolution and out of view. In terms of background clutters, motion blur, and fast motion, our method also achieves the second-best success rate. In conclusion, the proposed method is more stable and robust against different visual tracking challenges to a larger extent. Table 3 illustrates the average evaluation of all the ranked trackers, where the proposed algorithm obtains real-time value on a large number of short-term sequences with the same visual properties as the given dataset.

5. Conclusion

This paper proposes an optimal metric evaluation-based multicue inverse sparse appearance model for tracking the algorithm. According to the previous discussion, several key advantages exist in our method. Firstly, our scheme facilitates the reduction of redundant particle samples in each looping procedure with the help of the multicue inverse sparse appearance model in the global level. It not only explores the potential structural relationship among various ROI patches but also keeps the unbiasedness and dynamic stochasticity between consecutive frames. To dynamically depict the compact appearance model, crucial particle samples are extracted in the global level that alleviates the original particle filtering massive one-to-one matching computation effectively. This way also yields more precise representation for the target ROI whose effective numerical mounts of particle samples are limited to regularize the particle filtering process. Secondly, in order to select the most optimal particle sample among the previous crucial

cases, patch-based low-rank self-representation provides more robust and important training samples (atoms) for constructing the effective sparse coding book (dictionary) in the local level. Explicitly, it avoids opaque clustering for dictionary atoms which depend on manual initialization in a certain degree. Moreover, a set of structural pyramid-pooling process facilitates sparse coefficient-based optimal metric evaluation. In addition, iteratively templates update, and online metric training are included for updating the appearance model in the dynamic update process. Extensive evaluations on the test video sequences have demonstrated the effectiveness of the proposed method with favorable performance. Currently, we are working on a new algorithm that merges multitensors or deep features which are expected to save the computation cost and improve the robust tracking.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Leading Talents of Shandong University of Science and Technology, 863 Project: Physical Model-Based Dynamic Evolution Technology of Complex Scene (2015AA016404), Shandong Province Higher Educational Science and Technology Program (J17KA075), and the National Nature Science Foundation of China (61801270).

References

- [1] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: a benchmark," in *Proceedings of the 26th CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, Portland, Oregon, February 2013.
- [2] K. Li, F. He, H. Yu, and X. Chen, "A parallel and robust object tracking approach synthesizing adaptive Bayesian learning and improved incremental subspace learning," *Frontiers of Computer Science*, vol. 13, pp. 1116–1135, 2019.
- [3] X. Li, W. Hu, Z. Zhang, X. Zhang, and G. Luo, "Robust visual tracking based on incremental tensor subspace learning," in *IEEE 11th International Conference on Computer Vision, ICCV*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.
- [4] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.
- [5] H. Wang and T. Xu, "Robust visual tracking with incremental subspace learning sparse model," in *International Conference*

- in *Communications, Signal Processing, and Systems*, pp. 2721–2728, Springer, Harbin, China, July 2017.
- [6] G. Yang, Z. Hu, and J. Tang, “Robust visual tracking via incremental subspace learning and local sparse representation,” *Arabian Journal for Science and Engineering*, vol. 43, no. 2, pp. 627–636, 2018.
 - [7] W. Zhong, H. Lu, and M. Yang, “Robust object tracking via sparsity-based collaborative model,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1838–1845, Providence, Rhode Island, June 2012.
 - [8] X. Jia, H. Lu, and M. Yang, “Visual tracking via adaptive structural local sparse appearance model,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1822–1829, Providence, Rhode Island, June 2012.
 - [9] H. Kashiyani and S. B. Shokouhi, “Patchwise object tracking via structural local sparse appearance model,” in *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 125–131, Mashhad, Iran, October 2017.
 - [10] M. Zhao, H. Qian, R. Ying-Jiao, and G. Chen, “Robust object tracking via sparse representation based on compressive collaborative Haar-like feature space,” in *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 274–278, Shanghai, China, July 2016.
 - [11] G. Han, H. Luo, J. Liu, N. Sun, K. Du, and X. Li, “Multi-band joint local sparse tracking via wavelet transforms,” *IET Computer Vision*, vol. 10, no. 8, pp. 894–904, 2016.
 - [12] Y. Wu, M. Pei, M. Yang, Y. He, and Y. Jia, “Landmark-based inductive model for robust discriminative tracking,” in *Asian Conference on Computer Vision*, pp. 320–335, Singapore, November 2014.
 - [13] Y. Wu, M. Pei, M. Yang, J. Yuan, and Y. Jia, “Robust discriminative tracking via landmark-based label propagation,” *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1510–1523, 2015.
 - [14] Y. Wu, J. Wang, and H. Lu, “Real-time visual tracking via incremental covariance model update on log-euclidean riemannian manifold,” in *CCPR 2009 Chinese Conference on Pattern Recognition*, pp. 1–5, Nanjing, China, 2009.
 - [15] H. Grabner, M. Grabner, and H. Bischof, “Real-time tracking via on-line boosting,” in *Proceedings of the British Machine Vision Conference 2006 Bmvc*, vol. 1, p. 6, Edinburgh, UK, September 2006.
 - [16] C. Leistner, M. Godec, A. Saffari, and H. Bischof, “On-line multi-view forests for tracking,” in *DAGM Conference on Pattern Recognition*, pp. 493–502, Darmstadt, Germany, September 2010.
 - [17] T. Qin, B. Zhong, T. J. Chin, and H. Wang, “Matting-driven online learning of Hough forests for object tracking,” in *International Conference on Pattern Recognition*, pp. 2488–2491, Portland, Oregon, February 2013.
 - [18] Y. Zheng, L. Sun, S. Wang, J. Zhang, and J. Ning, “Spatially regularized structural support vector machine for robust visual tracking,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3024–3034, 2019.
 - [19] B. Babenko, M. H. Ming-Hsuan Yang, and S. Belongie, “Robust object tracking with online multiple instance learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.
 - [20] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, “Fully-convolutional siamese networks for object tracking,” in *European Conference on Computer Vision*, pp. 850–865, Springer, Amsterdam, Netherlands, October 2016.
 - [21] S. Hong, T. You, S. Kwak, and B. Han, “Online tracking by learning discriminative saliency map with convolutional neural network,” in *International Conference on Machine Learning*, pp. 597–606, Lille, France, July 2015.
 - [22] L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual tracking with fully convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3119–3127, Santiago, Chile, December 2015.
 - [23] K. Zhang, Q. Liu, Y. Wu, and M. H. Yang, “Robust visual tracking via convolutional networks without training,” *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1779–1792, 2016.
 - [24] L. Wang, W. Ouyang, X. Wang, and H. Lu, “STCT: sequentially training convolutional networks for visual tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1373–1381, Las Vegas, NV, USA, June 2016.
 - [25] D. Wang, H. Lu, Z. Xiao, and M. H. Yang, “Inverse sparse tracker with a locally weighted distance metric,” *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 24, no. 9, pp. 2646–2657, 2015.
 - [26] Y. Zhou, J. Han, X. Yuan, Z. Wei, and R. Hong, “Inverse sparse group lasso model for robust object tracking,” *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1798–1810, 2017.
 - [27] E. Elhamifar, G. Sapiro, and R. Vidal, “See all by looking at a few: sparse modeling for finding representative objects,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1600–1607, Providence, RI, USA, June 2012.
 - [28] S. Xiang, F. Nie, and C. Zhang, “Learning a Mahalanobis distance metric for data clustering and classification,” *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.
 - [29] D. Wang, H. Lu, and C. Bo, “Visual tracking via weighted local cosine similarity,” *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1838–1850, 2015.
 - [30] I. Markovsky, *Low Rank Approximation: Algorithms, Implementation, Applications*, Springer, NY, USA, 2011.
 - [31] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: spatial pyramid matching for recognizing natural scene categories,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2, pp. 2169–2178, IEEE, NY, USA, June 2006.
 - [32] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
 - [33] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216, ACM, Corvallis, Oregon, June 2007.
 - [34] K. Nummiaro, E. Koller-Meier, and L. Van Gool, “An adaptive color-based particle filter,” *Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2003.
 - [35] J. Mairal, F. Bach, and J. Ponce, “Sparse modeling for image and vision processing,” *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.
 - [36] H. Wang and H. Ge, “Object tracking via inverse sparse representation and convolutional networks,” *Optik - International Journal for Light and Electron Optics*, vol. 138, pp. 68–79, 2017.
 - [37] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, “Learning spatially regularized correlation filters for visual tracking,” *CoRR Clinical Orthopaedics and Related Research*, pp. 4310–4218, 2016.

- [38] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., vol. 8926, pp. 254–265, Springer, Zurich, Switzerland, September 2014, Lecture Notes in Computer Science.
- [39] Y. Qi, S. Zhang, L. Qin et al., "Hedged deep tracking," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 4303–4311, IEEE Computer Society, Las Vegas, NV, USA, June 2016.
- [40] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, BMVC*, M. F. Valstar, A. P. French, and T. P. Pridmore, Eds., BMVA Press, Nottingham, UK, September 2014.
- [41] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.
- [42] X. Mei and H. Ling, "Robust visual tracking using ℓ_1 minimization," in *IEEE 12th International Conference on Computer Vision, ICCV 2009*, pp. 1436–1443, IEEE Computer Society, Kyoto, Japan, September 2009.
- [43] C.-Y. Tsai and Y.-C. Feng, "Real-time multi-scale parallel compressive tracking," *Journal of Real-Time Image Processing*, vol. 16, no. 6, pp. 2073–2091, 2019.