Learning Discriminative Hidden Structural Parts for Visual Tracking

Longyin Wen[1](✉), Zhaowei Cai[1], Dawei Du[2], Zhen Lei[1], and Stan Z. Li[1]

[1] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China
{lywen,zwcai,zlei,szli}@nlpr.ia.ac.cn

[2] School of Computer and Control Engineering, University of Chinese
Academy of Sciences, No.3, Zhongguancun Nanyitiao, Beijing 100049, China
dawei.du@vipl.ict.ac.cn
http://www.cbsr.ia.ac.cn

Abstract. Part-based visual tracking is attractive in recent years due to its robustness to occlusion and non-rigid motion. However, how to automatically generate the discriminative structural parts and consider their interactions jointly to construct a more robust tracker still remains unsolved. This paper proposes a discriminative structural part learning method while integrating the structure information, to address the visual tracking problem. Particulary, the state (e.g. position, width and height) of each part is regarded as a hidden variable and inferred automatically by considering the inner structure information of the target and the appearance difference between the target and the background. The inner structure information considering the relationship between neighboring parts, is integrated using a graph model based on a dynamically constructed pair-wise Markov Random Field. Finally, we adopt Metropolis-Hastings algorithm integrated with the online Support Vector Machine to complete the hidden variable inference task. The experimental results on various challenging sequences demonstrate the favorable performance of the proposed tracker over the state-of-the-art ones.

1 Introduction

Visual tracking is one of the most important and challenging problems in computer vision field. Traditionally, the majority of existing methods always focus on modeling the holistic appearance of the target within a bounding box, and they have achieved good performance in some conditions [1–8]. Intuitively however, they ignore the local information of the target, which greatly limits their application in the scenario where the target is partially occluded or the global appearance of the target changes a lot.

Recently, methods [9–18] with part-based appearance representation are popular in visual tracking task to deal with some situations that the holistic appearance based methods fail. The part based tracker combines multiple parts with local information to achieve stronger representation ability, and is able to explicitly model the target structure variations. However, the obvious shortcomings

C.V. Jawahar and S. Shan (Eds.): ACCV 2014 Workshops, Part III, LNCS 9010, pp. 262–276, 2015.
DOI: 10.1007/978-3-319-16634-6_20

of these previous part based trackers are: (1) the number of parts are assigned before empirically, which is hard to obtain better discriminative ability with suitable parts; (2) no relationships between neighboring generated parts are considered, which loses the structural information of the target. Therefore, their methods lack strong discriminative ability and fail to combine the parts into a whole to complete the tracking task.

In this paper, we propose a discriminative Hidden Structural Part Tracker (HSPT), which tracks arbitrary objects without any assumptions on the scenarios. The proposed method learns the discriminative parts automatically by integrating the structure and discriminative information. In the learning step, both the appearance of the parts and the relationships between them are considered. Since the discriminative parts are not located in the fixed location to the target center, we regard the state of them as hidden variables in the objective function. Then, the objective is optimized by the Metropolis-Hastings (MH) algorithm [19] integrated with the online Support Vector Machine (SVM) method [20] iteratively. In order to achieve more robust performance in complex environments, the bounding box based appearance of the target is also incorporated in our tracker. The contributions of this paper are concluded as follows:

- We propose a hidden discriminative structural parts learning based tracker, which simultaneously learns multiple structural parts of the target to represent the target better to enhance its robustness.
- The MH algorithm and the online SVM are interestingly combined to infer the optimal state of the discriminative parts, and this optimization method handles varying number of parts well.
- The structural supporting between parts are naturally integrated through the dynamically constructed pair-wise MRF model.
- Extensive tracking experiments are various publicly available challenge sequences demonstrate the favorable performance against the state-of-the-art methods.

2 Related Works

The part based model has been developed recently in visual tracking task. Shahed et al. proposed a part-based tracker HABT [10], which generated the target parts by manually labeling in the first frame. And the appearance model of the target is assumed to be fixed in the tracker, which limits its performance in the complex environment. Another tracker Frag [9] utilized the regularly partitioned parts to model the appearance of the target, but the appearance model of each part is also fixed. In [18], Yao et al. introduced an online part-based tracking with latent structured learning. However, the proposed method uses fix number of parts to represent the target object, which makes it insufficient to describe the appearance variations when large deformation happens.

There also exist some other part based appearance representation methods. Kwon et al. [11] proposed the BHMC tracker, which generates the parts based on SURF-like key points without global structure constraints and the appearance of the parts were updated roughly. Wang et al. [14] proposed a discriminative appearance model based on superpixels, called SPT, in which the probabilities of superpixels belonging to the foreground were utilized to discriminate the target from the background. However, the relation between superpixels was not incorporated, which makes the tracker easily affected by the similar backgrounds. Godec et al. [12] extended the hough forest to the online domain and integrated the voting method for tracking, regardless of the structure information. Cehovin et al. [13] proposed a coupled-layer visual tracker, which combined the global and local appearance of the target together in part generation. However, the ignorance of the relationships between parts makes it unstable when clutter backgrounds, occlusions or non-rigid motion happen. Cai et al. [17] designed a dynamic graph based method which works well in non-rigid motion, but the parts are generated only based on color feature and some background parts will be easily misclassified as foreground.

3 Overview of Proposed Method

Generally, the tracking task is formulated as a Markovian state transition process, where the current target state is determined by its previous state. Let $p(O_t|Z_t)$ be the appearance model and $p(Z_t|Z_{t-1})$ be the motion model of the target at time t. The state of the target is represented as $Z_t = (\ell_t, s_t)$. ℓ_t is the position of the target in the 2D image plane and s_t is the size of the target consisting of the width and height. The motion model $p(Z_t|Z_{t-1})$ and the appearance model $p(O_t|Z_t)$ are described as follows.

Motion Model. Similar to [2], we assume the position and size of the target varies independently in the motion model, that is:

$$p(Z_t|Z_{t-1}) = p(\ell_t, s_t|\ell_{t-1}, s_{t-1}) = p(\ell_t|\ell_{t-1})p(s_t|s_{t-1}), \tag{1}$$

where the target position transition probability $p(\ell_t|\ell_{t-1}) = 1$, if $\|\ell_t - \ell_{t-1}\|_2 < R_s$; otherwise, it equals to zero. R_s is the predefined searching radius. The target scale transition probability is similarly handled as [2].

Appearance Model. The appearance of our tracker consists of two parts, the learned discriminative structural parts model $\mathcal{A}^{(0)}$ and the bounding box based appearance model $\mathcal{A}^{(1)}$. The learned structural parts focus on the local variations and the bounding box based appearance focuses on the holistic variations of the target. Intuitively, the combination of them can achieve more robust performance. The appearance model of our tracker is formulated as follows:

$$p(O_t|Z_t) = \Big(p^{(0)}(O_t|Z_t)\Big)^{\lambda_b} \cdot \Big(p^{(1)}(O_t|Z_t)\Big)^{(1-\lambda_b)}, \tag{2}$$

where λ_b is a predefined balance parameter, and $p^{(0)}(O_t|Z_t)$ and $p^{(1)}(O_t|Z_t)$ are the probabilities of the target candidate given out by $\mathcal{A}^{(0)}$ and $\mathcal{A}^{(1)}$, respectively. To model the appearance of the target, the online SVM [20] is adopted for each part.

For the discriminative structural parts model $\mathcal{A}^{(0)}$, the probability of the candidate, including the appearance likelihood and the deformation likelihood of the parts, is calculated as

$$p^{(0)}(O_t|Z_t) = \prod_{i,j \in \mathcal{N}} \phi(Z_{i,t}, Z_{j,t}) \cdot \prod_{i=1}^{n} p(O_{i,t}|Z_{i,t}), \tag{3}$$

where n is the number of parts, $p(O_{i,t}|Z_{i,t})$ is the probability of the part i applauding the candidate to be positive, $O_{i,t}$ is the observation of part i, $\mathcal{N}$ is the neighboring system of the parts and $\phi(Z_{i,t}, Z_{j,t})$ is the pairwise interaction potentials between the learned part i and part j.

In our model, the state of part i at time t is defined as $Z_{i,t} = (x_{i,t}, y_{i,t}, w_{i,t}, h_{i,t}, \varDelta x_{i,t}, \varDelta y_{i,t})$, where $(x_{i,t}, y_{i,t})$, $w_{i,t}$ and $h_{i,t}$ are the position, width and height of part i at time t respectively. $(\varDelta x_{i,t}, \varDelta y_{i,t})$ is the spatial offset of the part relative to the target center. The interaction potential term is expressed by means of Gibbs distribution:

$$\phi(Z_{i,t}, Z_{j,t}) \propto \exp\Big(- \lambda_{\phi} \|v_t(i,j) - \tilde{v}(i,j)\|_2 \Big), \tag{4}$$

where $\lambda_{\phi} = 0.2$ is the scaler parameter in the experiments and $v_t(i,j) = \ell_t(i) - \ell_t(j)$ represents the vector pointing from the position $\ell_t(i)$ of part i to the location $\ell_t(j)$ of part j at time t, which encodes the supporting between neighboring parts (structural information of the target). $\tilde{v}(i,j)$ is the learnt relative position of the part i and j. Here, we model the relationship between different parts, rather than modeling the exclusions between close targets as in [21].

Let $\varPhi_p(O_{i,t})$ represent the HOG feature [22] of the part observation, and $\omega_{i,p}^{t}$ is the SVM parameter corresponding to part i at time t. The likelihood of its appearance is calculated as

$$p(O_{i,t}|Z_{i,t}) \propto \exp\big(\omega_{i,p}^{(t)} \cdot \varPhi_p(O_{i,t})\big), \tag{5}$$

In order to reduce the influence of some badly learned parts, we only utilize η percent high confident parts to score the candidates. Then the probability of the candidate (3) can be rewritten as

$$p^{(0)}(O_t|Z_t) \propto \prod_{i,j \in \mathcal{N}} \phi(Z_{i,t}, Z_{j,t}) \cdot \exp\Big(\sum_{i \in \mathcal{I}} \omega_{i,p}^{(t)} \cdot \varPhi_p(O_{i,t}) \Big), \tag{6}$$

where $\mathcal{I}$ is the index set of the selected η high confident parts.

For the bounding box based appearance model $\mathcal{A}^{(1)}$, the probability of the candidate is presented as:

$$p^{(1)}(O_t|Z_t) \propto \exp(\omega_b^{(t)} \cdot \varPhi_b(O_t)), \tag{7}$$

where the SVM parameter $\omega_b^{(t)}$ and the HOG feature $\varPhi_b(O_t)$ are determined based on the whole target.

4 Learning the Discriminative Parts

In the tracking task, the appearance of the target changes dramatically. In order to adapt the part models to the target appearance variations, some of the target parts should be added or deleted, and their state should be determined to represent the target optimally. Therefore, we design a reasonable objective function to perform the parts learning, whose goal consists of three aspects: (1) maximize the margin between the target and the background; (2) retain the structure information of the target; (3) cover the most of the target foreground area. The state of each part is treated as a hidden variable, which will be inferred based on the acquired observation information.

As discussed above, our objective in terms of optimization is to find the optimized SVM parameter $\omega_{i,p}$ and parts state $Z_{i,t}$. In this section, the MH algorithm and the online SVM are integrated into an unified optimization framework to complete the inference task. The details will be presented as follows.

4.1 Objective

For the local parts to be learned, we expect that they acquire better representation ability to ensure robust tracking performance. The objective for the i-th part learning is formulated as

$$G(\omega_{i,p}, Z_i; \mathcal{X}) = \alpha \cdot \rho \cdot \omega_{i,p} \cdot \varPhi_p(\mathcal{X}_{Z_i}) + \beta \cdot \mathcal{R}(\mathcal{F}, Z_i), \tag{8}$$

where the first term is to separate the target parts from the background parts by maximizing the margin between them, and the second term encourages the learned part to cover more target foreground area. $\mathcal{F}$ is foreground area, $\rho \in \{-1, 1\}$ is the binary label of the updating sample $\mathcal{X}$ indicating the foreground and background, and $\mathcal{X}_{Z_i}$ represents the part observations of the updating sample with the part state Z_i. $\mathcal{R}(\mathcal{F}, Z_i)$ means the overlap ratio between Z_i and $\mathcal{F}$. We set the balancing parameters $\alpha = 0.7$, $\beta = 0.2$ in all of our experiments.

Naturally, we infer the optimal state of each part jointly and integrate the structure information in the inference process. The objective for the target is proposed as

$$G(\omega_p, \tilde{Z}; \mathcal{X}) = \alpha \cdot \rho \cdot \omega_p \cdot \varPhi_p(\mathcal{X}_{\tilde{Z}}) + \beta \cdot \mathcal{R}(\mathcal{F}, \tilde{Z}), \tag{9}$$

where $\tilde{Z} = (Z_1, \cdots, Z_n)$ is the combination of the parts state and $\omega_p = (\omega_{1,p}^{(t)}, \cdots, \omega_{n,p}^{(t)})$ is the concatenated SVM weight of each part. $\mathcal{R}(\mathcal{F}, \tilde{Z}) = \sum_{i=1}^{n} \mathcal{R}(\mathcal{F}, Z_i)$ represents the coverage ratio of the true target area.

4.2 Optimal Parts Inference

The MH algorithm [19] has been applied in the multiple target tracking task [21,23], where the authors use it for particle filter sampling to identify the state of each target precisely. However, in our paper, we do not aim at identifying each part all the time. Instead, we utilize the MH algorithm to clean up useless parts and discover new parts adaptively according to the target appearance variation. Firstly, we need convert our objective (9) into the probability form:

$$p(\tilde{Z}, \omega_p|\mathcal{X}) \propto \exp\Big(\zeta \cdot G(\omega_p, \tilde{Z}; \mathcal{X})\Big), \tag{10}$$

where ζ is the scale factor. Then maximizing the objective (9) is equivalent to solve the maximum posterior probability problem:

$$\{\tilde{Z}, \omega_p\} = \arg\max_{\tilde{Z},\omega_p} p(\tilde{Z}, \omega_p|\mathcal{X}). \tag{11}$$

Due to the dependance between the hidden variable $\tilde{Z}$ and ω_p, it is difficult to optimize them simultaneously. Hence, we decompose the inference task of the two hidden variables in (10) into a two stage iterative optimization problem. In each pass r, we solve the objective by dividing it into two steps to iteratively update $\{\tilde{Z}, \omega_p\}$ using the following procedure.

Optimize $\boldsymbol{\omega_p}$. Given the optimized parts state $\tilde{Z}^{(r)}$, (11) is equivalent to the following optimization problem:

$$\omega_p^{(r)} = \arg\max_{\omega_p} p(\tilde{Z}^{(r)}, \omega_p|\mathcal{X}) = \arg\max_{\omega_p} \Big\{\alpha \cdot \rho \cdot \omega_p \cdot \varPhi_p(\mathcal{X}_{\tilde{Z}^{(r)}})\Big\}. \tag{12}$$

Then in analogy to classical SVM, we train the parameter $\omega_p^{(r)}$ by solving the following optimization problem:

$$\omega_p^{(r)} = \arg\min_{\omega_p} \Big\{\frac{1}{2}\|\omega_p\|^2 + \gamma \sum_{i=1}^{m} \max\big(0, 1 - \rho^{(i)} \cdot f_{\omega_p}(\mathcal{X}^i)\big)\Big\}, \tag{13}$$

where $\{(\mathcal{X}^1, \rho^1), \cdots, (\mathcal{X}^m, \rho^m)\}$ is the collected sample pool, $\rho^i \in \{-1, 1\}$ is the label of the i^{th} collected sample $\mathcal{X}^i$, m is the number of samples. We set parameter $\gamma = 5$ in our experiments. The score of the sample is calculated as $f_{\omega_p}(\mathcal{X}_{\tilde{Z}^{(r)}}) = \omega_p \cdot \varPhi_p(\mathcal{X}_{\tilde{Z}^{(r)}})$, where $\varPhi_p(\mathcal{X}_{\tilde{Z}^{(r)}})$ is the concatenated HOG feature of parts.

Optimize $\boldsymbol{\tilde{Z}}$. With the determined $\omega_p^{(r)}$, we sample a proposal part state $\tilde{Z}^{(r)\prime}$ according to the previous parts state $\tilde{Z}^{(r)}$, and calculate the acceptance ratio in the MH algorithm based on the optimized model parameter $\omega_p^{(r)}$ to get the optimized parts state $\tilde{Z}^{(r+1)}$. Therefore, the MAP solution of the parts state in the constructed Markov Chain of MH algorithm is utilized to get the optimized state $\tilde{Z}$.

To that end, five moves are defined for states change of each part. *Birth*, indicates the move of adding the candidate parts in the sampler; *Death*, indicates the move of removing the newly added candidate parts in the sampler. The reversible pair focuses on the newly added candidates generated by SLIC sampling [24] (i.e. the external rectangle region of the generated superpixel is used as the candidate) and the deleted candidates, if the target is changing the pose so that some old parts will disappear and some new parts will be generated. *Stay*, indicates the move of adding the disappeared parts in previous iterations in the sampler; *Leave*, indicates the move of removing the learned parts in previous iterations in the sampler. The reversible pair determine the state of the learned parts in previous sampling iterations when the target undergoes heavy occlusion so that some old parts are missed temporarily and appear again then. *Update*, indicates the move of updating the parts in the sampler, which deals with the dynamically updated appearance of the target as a self-reversible pair.

For easy description, we omit the iteration mark r in the following. We define two sets in the optimization process: (1) $T^\star = \{T_1^\star, \cdots, T_n^\star\}$ is the learned parts set and its corresponding state set is $Z^\star = \{Z_1^\star, \cdots, Z_n^\star\}$; (2) $T^+ = \{T_1^+, \cdots, T_m^+\}$ is the birth candidate set and its corresponding state set is $Z^+ = \{Z_1^+, \cdots, Z_m^+\}$. The notation $Z_i^{(\cdot)}$ ($Z_i^\star$ or Z_i^+) is the current part state and $Z_i^{(\cdot)\prime}$ is the proposal state of the part.

Let $\mathcal{N}_{i,t}$ be the neighbors of part i at time t. Let $C = \{C_b, C_d, C_s, C_l, C_u\}$ represent the prior probability of each move type and we set $C = \{0.3, 0.1, 0.1, 0.01, 3.0\}$ in the experiments empirically. The proposal distribution $q = \{q_b, q_d, q_s, q_l, q_u\}$ and the acceptance ratio are calculated as follows.

Birth: Select the part T_i^+ from the birth candidate parts set T^+ with the uniform distribution. The birth proposal distribution can be calculated as $q_b(Z_i^{+\prime}; Z_i^+) = \frac{C_b}{m}$, if $(T^{\star\prime}, Z^{\star\prime}) = (T^\star \cup \{T_i^+\}, Z^\star \cup \{Z_i^+\})$, and otherwise it equals to zero. Then the acceptance ratio is presented as

$$\alpha_b = \min\left(1, p(\mathcal{X}|Z_i^{+\prime}, \omega_p) \cdot p(Z_i^{+\prime}) \cdot \frac{q_d(Z_i^+; Z_i^{+\prime})}{q_b(Z_i^{+\prime}; Z_i^+)}\right), \tag{14}$$

where $p(Z_i^{+\prime})$ represents the birth transition probability.

Death: Select the part $T_i^\star$ as the death part with the uniform distribution from the newly added candidate set $T^+ \cap T^\star$. The death proposal distribution is defined as $q_d(Z_i^{\star\prime}; Z_i^\star) = \frac{C_d}{|T^\star \cap T^+|}$, if $(T^{\star\prime}, Z^{\star\prime}) = (T^\star \backslash \{T_i^+\}, Z^\star \backslash \{Z_i^+\})$, and otherwise it equals to zero. Then the acceptance ratio is presented as

$$\alpha_d = \min\left(1, \frac{1}{p(\mathcal{X}|Z_i^{\star\prime}, \omega_p)} \cdot \frac{1}{p(Z_i^{\star\prime})} \cdot \frac{q_b(Z_i^\star; Z_i^{\star\prime})}{q_d(Z_i^{\star\prime}; Z_i^\star)}\right), \tag{15}$$

Stay: Select the part $T_i^\star$ to be the stay part. We introduce a set $T^{(d)} = \mathcal{T}^\star \backslash T_i^\star$, $\mathcal{T}^\star$ is the union set of the parts set in the previous iterations. Then the stay proposal distribution is presented as $q_s(Z_i^{\star\prime}; Z_i^\star) = \frac{C_s}{|T^{(d)}|} \cdot J(Z_i^{\star\prime})$, if $|T^{(d)}| \neq 0$, and otherwise it equals to zero. The function $J(Z_i^{\star\prime})$ represents the probability of part $T_i^\star$ staying at the state $Z_i^{\star\prime}$, and it is modeled as a normal density centered at the disappearing point. Z_i^l is the state of the part $T_i^\star$ at the disappearing point in the previous iterations. Then the acceptance ratio is presented as

$$\alpha_s = \min\left(1, p(\mathcal{X}|Z_i^{\star\prime}, \omega_p) \cdot \prod_{j \in \mathcal{N}_t} \phi(Z_i^{\star\prime}, Z_{j,t}) \cdot p(Z_i^{\star\prime}|Z_i^\star) \cdot \frac{q_l(Z_i^\star; Z_i^{\star\prime})}{q_s(Z_i^{\star\prime}; Z_i^\star)}\right), \tag{16}$$

where $\mathcal{N}_t$ is the part neighboring system at t, $p(Z_i^{\star\prime}|Z_i^l)$ is the state transition probability.

Leave: Select the part $T_i^\star$ to be the leave part. The leave proposal distribution is defined as $q_l(Z_i^{\star\prime}; Z_i^\star) = \frac{C_l}{|T^\star|}$, if $(T^{\star\prime}, Z^{\star\prime}) = (T^\star \backslash \{T_i^\star\}, Z^\star \backslash \{Z_i^\star\})$. Otherwise, it equals to zero. Then the acceptance ratio is presented as

$$\alpha_l = \min\left(1, \frac{1}{p(\mathcal{X}|Z_i^{\star\prime}, \omega_p)} \cdot \frac{1}{\prod_{j\in\mathcal{N}_t} \phi(Z_i^{\star\prime}, Z_{j,t})} \cdot \frac{1}{p(Z_i^{\star\prime}|Z_i^\star)} \cdot \frac{q_s(Z_i^\star; Z_i^{\star\prime})}{q_l(Z_i^{\star\prime}; Z_i^\star)}\right), \quad (17)$$

Update: Select the part $T_i^\star$ to be the update part. The appearance of the part will be updated if the move is accepted. The proposal distribution of this move type is defined as $q_u(Z_i^{\star\prime}; Z_i^\star) = \frac{1}{|T^\star|}$, if $T_i^\star \in T^\star$. Otherwise it equals to zero. Then the acceptance ratio is presented as

$$\alpha_u = \min\left(1, \frac{p(\mathcal{X}|Z_i^{\star\prime}, \boldsymbol{\omega}_p)}{p(\mathcal{X}|Z_i^\star, \boldsymbol{\omega}_p)} \cdot \frac{\prod_{j\in\mathcal{N}_t} \phi(Z_i^{\star\prime}, Z_{j,t})}{\prod_{j\in\mathcal{N}_t} \phi(Z_i^\star, Z_{j,t})}\right), \quad (18)$$

In the above acceptance ratio calculation Eqs. (14–18), the birth transition probability $p(Z_i^{(\cdot)}) = \exp(-\lambda_o \cdot \mathcal{R}(Z_i^{(\cdot)}, Z_i^\star))$. This term penalizes the overlap ratio between the candidate part T_i^+ and the existing learned parts to avoid adding redundant parts. We set $\lambda_o = 2$ in our experiments. The probability $p(\mathcal{X}|Z_i^{(\cdot)}, \omega_p) \propto \exp(G(\mathcal{X}; \omega_p, Z_i^{(\cdot)}))$, where $\mathcal{X}$ is the current updating sample and $Z_i^{(\cdot)}$ ($Z_i^\star$, $Z_i^{\star\prime}$, Z_i^+ or $Z_i^{+\prime}$) is the part state.

In addition, the target parts interact with each other, especially for the neighboring ones, so it is inappropriate to assume the independence between target parts. We should integrate the relationships between parts in optimization process rather than optimize the objective in (8) for each part individually. Motivated by [21], we propose to utilize the dynamically constructed pairwise MRF to model the supporting between different parts in the part learning process. We set all pairs of parts as the neighbors in the graph. Thus, the part transition probability in (16) and (17) is presented as follows:

$$p(\tilde{Z}'|\tilde{Z}) \propto \prod_i p(Z_i'|Z_i) \prod_{ij\in\mathcal{N}} \phi(Z_i', Z_j'), \quad (19)$$

where $\tilde{Z} = (Z_1, \cdots, Z_n)$ is the combination state of multiple parts, $\phi(Z_i', Z_j')$ is the pairwise interaction potentials between parts similarly defined as (4), $p(Z_i'|Z_i)$ is the part transition model, and $\mathcal{N}$ is the neighbor system of the parts. Z_i' is the proposal state of part i and Z_i is the state of part i currently. The part transition is modeled as a normal density centered at the previous state, which is presented as $Z_i'|Z_i = Z_i + \Delta Z_i$, where $\Delta Z_i \sim [\mathcal{N}(0, \sigma_x^2), \mathcal{N}(0, \sigma_y^2), \mathcal{N}(0, \sigma_w^2), \mathcal{N}(0, \sigma_h^2), \mathcal{N}(0, \sigma_{\Delta x}^2), \mathcal{N}(0, \sigma_{\Delta y}^2)]$.

Figure 1 is an example to illustrate how the discriminative parts are automatically learned over time in *shirt* sequence. The shirt is crinkled in frame ♯0026 and ♯0040, where the target bounding box contains considerable background in

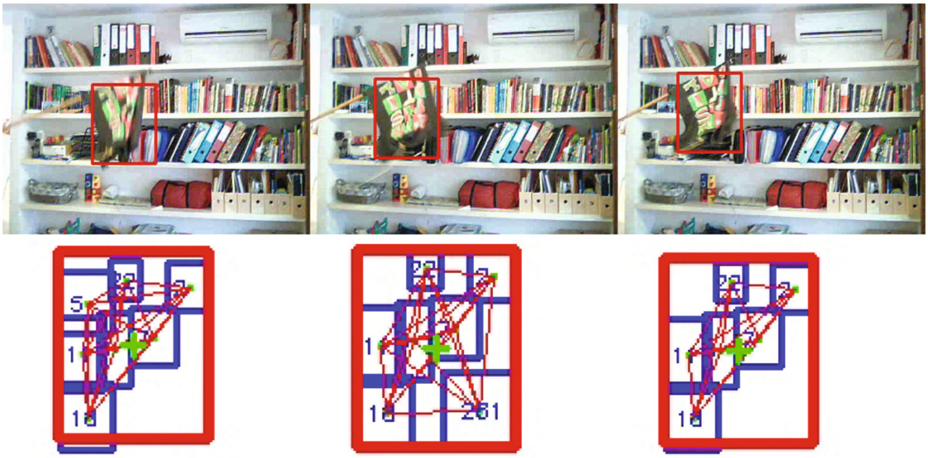

Fig. 1. The first row is the tracking results of our tracker in the sequence *shirt* and the second row presents the learned discriminative parts and the corresponding structure. The nodes in the graph represent each learned part and the lines represent the spatial relationships between neighboring parts. The green cross represents the center of the target (Colour figure online).

the right bottom corner. In contrast, the target bounding box in frame ♯0038 contains only foreground. In this case, the proposed part-based method adaptively generates a part 261 to cover the new foreground in frame ♯0038, and deletes it in frame ♯0040. In this way, our part based model can adapt to the variations of the target better than the bounding box based model. The final optimization scheme is summarized in Algorithm 1.

5 Experimental Results

5.1 Parameters

The parameters in our experiment are detailed in the following. In the learning phase, we run $P_n = 400$ iterations to complete the part state inference task and $P_{n_0} = 100$ of them are burn-in in the MH algorithm. Generally, the algorithm will converge after about 300 iterations. The motion model parameters we used are $\sigma_x^2 = 3$, $\sigma_y^2 = 3$, $\sigma_w^2 = 0.2$, $\sigma_h^2 = 0.2$, $\sigma_{\Delta x}^2 = 0.1$ and $\sigma_{\Delta y}^2 = 0.1$. The cell size of HOG is set as 8×8 pixels. A block consists of 4 cells and the strides of the cell are set 4 in both x and y directions. The linear kernel is exploited in the online learning SVM model. The target area is divided into about 15 or 20 superpixels in the SLIC algorithm. Meanwhile, in the tracking phase, the searching radius $R_s \in [20, 60]$. The balance parameter λ_b in (2) is set in the interval $[0, 0.5]$.

Algorithm 1. Discriminative Hidden Structural Part Tracker

```
Initialize the target state $\hat{Z}_1$.
for $t = 2$ to $N$ do
    Get the foreground $\mathcal{F}$ and collect the birth candidate parts based on the optimized target state $\hat{Z}_t = \hat{Z}_{t-1}$, and get the initial joint parts state $\tilde{Z}^{(1)}$.
    Set the sample set $\Gamma = \emptyset$ in the MH algorithm.
    for $r = 1$ to $P_n$ do
        Get optimal $\omega_p^{(r)}$ based on the current state $\tilde{Z}^{(r)}$.
        Generate the proposal joint state $\tilde{Z}^{(r)\prime}$ based on $\tilde{Z}^{(r)}$:
          – Choose a move type according to the move prior probability $C$.
          – Select a part $i$ according to the move proposal distribution $q$ and compute the acceptance ratio $\alpha$ for $\tilde{Z}^{(r)\prime}$ in (14), (15), (16), (17), (18).
          – Accept the proposal state, if $\alpha \geq 1$, and add it to $\Gamma$, $\tilde{Z}^{(r+1)} = \tilde{Z}^{(r)\prime}$; otherwise generate a uniform random number $u \in [0, 1]$. If $u < \alpha$, accept the proposal state and add it to $\Gamma$, $\tilde{Z}^{(r+1)} = \tilde{Z}^{(r)\prime}$; otherwise reject the proposal state and add the previous state to $\Gamma$, $\tilde{Z}^{(r+1)} = \tilde{Z}^{(r)}$.
          – $\tilde{Z}^{(r)} = \tilde{Z}^{(r+1)}$.
    end for
    Discard the first $P_{n_0}$ burn-in samples in $\Gamma$.
    Get the optimized parts $Z^*$ by the $MAP$ solution of $\Gamma$ in (11).
    Update the SVM model and the MRF graph model of the parts with $\{Z^*, \omega_p\}$ in $\mathcal{A}^{(0)}$, and update the SVM model with $\{\hat{Z}_t, \omega_b\}$ in $\mathcal{A}^{(1)}$.
end for
```

5.2 Effectiveness of $\mathcal{A}^{(0)}$ and $\mathcal{A}^{(1)}$

Firstly, we chose four representative sequences to demonstrate the behavior of $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(0)}$. The results shown in Table 1 demonstrate the performance of HSPT is improved mainly due to the local discriminative parts learning rather than the features or the classifiers adopted.

Table 1. Comparison results of $\mathcal{A}^{(1)}$, $\mathcal{A}^{(0)}$, and HSPT.

Seq.	AECP Metric			PASCAL VOC Metric		
	$\mathcal{A}^{(1)}$	$\mathcal{A}^{(0)}$	HSPT	$\mathcal{A}^{(1)}$	$\mathcal{A}^{(0)}$	HSPT
tiger2	26.7	**14.7**	**8.39**	80	**215**	**280**
shirt	30.2	**22.5**	**8.34**	680	**920**	**1310**
pedestrian	105	**14.9**	**3.76**	110	**310**	**355**
car	29.7	**10.3**	**7.78**	770	**870**	**895**

As presented in Table 1, the combined HSPT outperforms the individual $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(0)}$ in all tested sequences. $\mathcal{A}^{(1)}$ focuses on holistic appearance and $\mathcal{A}^{(0)}$ focuses on inner structure of the target. $\mathcal{A}^{(0)}$ is superior over $\mathcal{A}^{(1)}$ because the local discriminative structure model represents the target better than appearance only model. Especially in the sequences *shirt* and *pedestrian* where nonrigid deformations and illumination variations frequently happen, the targets were still well tracked by $\mathcal{A}^{(0)}$ even when several parts undergo changes in location and appearance. In contrast, $\mathcal{A}^{(1)}$ was affected more seriously by these challenges. $\mathcal{A}^{(1)}$ focuses on the holistic appearance, which can enhance the stability of the tracker in the complex situations. HSPT inherits the advantages both from $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(0)}$ and thus performed best against the other evaluated trackers on the evaluated sequences.

5.3 Comparison with Other Trackers

Then, we compare our tracker with some state-of-the-art methods, including bounding box based methods (MIL [2], VTD [7], $\ell1$ [4], TLD [5]), and part based methods (Frag [9], HABT [10], BHMC [11], SPT [14]). All the codes are provided by the authors on their websites. Ten challenging sequences (nine of them are publicly available [2,5,7,25,26] and the other one is collected by ourself) are utilized in the experiment. These sequences cover most of the challenging situations in tracking task: non-rigid motion, in-plane and out-of-plane rotation, large illumination changes, heavy occlusions and complex background (see Fig. 3).

Table 2. Comparison results based on the AECP metric.

Seqences	MIL	$\ell1$	TLD	VTD	Frag	HABT	BHMC	SPT	HSPT
football	12.7	26.3	13.1	**6.25**	9.92	70.0	70.9	170	**4.39**
tiger1	**8.35**	60.4	25.9	22.3	29.3	16.9	-	16.3	**8.29**
tiger2	**5.91**	47.3	25.3	31.5	39.3	53.7	-	-	**8.39**
david	15.6	77.3	**4.49**	31.7	55.9	44.7	-	144	**5.44**
shirt	32.1	68.4	70.8	27.5	19.9	79.5	**9.47**	36.6	**8.34**
pedestrian	64.3	128	36.7	86.0	56.9	80.8	-	27.1	**3.76**
stone	**9.07**	11.6	9.69	26.1	92.9	127	40.3	50.3	**6.14**
carchase	39.9	8.92	**3.77**	81.1	12.6	21.0	-	153	**3.95**
car	80.3	23.3	**11.1**	51.8	28.6	26.3	-	153	**7.78**
portman	42.5	60.6	30.7	45.1	54.5	39.4	**23.3**	73.4	**17.6**

Table 3. The successfully tracked frames based on the PASCAL VOC metric.

Seqences	Frames	MIL	$\ell1$	TLD	VTD	Frag	HABT	BHMC	SPT	HSPT
football	**362**	272	206	272	**357**	302	206	55	30	**362**
tiger1	**354**	**279**	80	150	189	155	**209**	-	65	**279**
tiger2	**365**	**315**	15	60	85	35	25	-	-	**280**
david	**462**	328	124	**422**	65	159	253	-	30	**462**
shirt	**1365**	760	55	5	760	**890**	10	330	595	**1310**
pedestrian	**355**	55	85	75	**110**	80	55	-	85	**355**
stone	**593**	135	**384**	339	**384**	95	5	40	115	**473**
carchase	**424**	91	283	**419**	66	368	318	-	61	**424**
car	**945**	105	765	**880**	575	650	480	-	25	**895**
portman	**301**	94	69	118	**168**	74	148	30	74	**271**

The proposed tracker is implemented in C++ and it runs about 0.1 fps on the Intel 3.0 GHz PC platform. We present the tracking results in this section and more results as well as demos can be found in supplementary materials. Two metrics are utilized to quantify the performance, namely the Average Error Center Location in Pixels (AECP) metric (↓) and the PASCAL VOC object detection metric [27] (↑). The symbol ↑ means methods with higher scores perform better, and ↓ indicate methods with lower scores perform better. The quantitative comparison results are shown in Tables 2 and 3.

Heavy Occlusion. The target in sequences *carchase*, *stone*, *car*, *tiger1* and *tiger2* undergoes heavy occlusion multiple times. As shown in Fig. 3, most of the trackers drift away when heavy occlusion happens, and some of them can not recapture the target after occlusion. In the sequence *car*, TLD with detection module works relatively well. Some other trackers such as Frag and $\ell 1$ who have intuitive robustness to occlusion also track the car well. However, our tracker still outperforms other methods due to the consideration of the relationships between parts which alleviates the influence of some badly learned parts.

Large Illumination Variations. The frequent large illumination variations in *tiger1*, *tiger2* and *pedestrian* sequences challenge the performance of the trackers. Since the appearance features are easily affected by illumination variations, most of the previously proposed trackers fail to track the target in these sequences. For example, when the light is shining in the sequence *tiger1* and *tiger2*, Frag fails to track the tiger, and when the woman is under the shadow in the *pedestrian* sequence, VTD and SPT shrink to those parts that are not shadowed. In contrast, our tracker optimally partitions the target into several parts, and the target can be located with the help of those less affected parts. The combined appearance features of different parts, the structure information between parts, and the structure information between the part and the target center make our tracker outperforms other trackers.

Pose Changes. The inner structure changes caused by pose changes usually make the bounding box based trackers fail to track the target. Nevertheless, part based trackers including Frag and BHMC work relatively well because they focus on the parts instead of the holistic bounding box template, which can be demonstrated in the sequence *portman* and *shirt* in Fig. 3, Tables 2 and 3. Especially in the sequence *shirt*, the bounding box based trackers such as $\ell 1$, VTD and MIL fail because of the error accumulation when the non-rigid motion happens. Since the combination of part appearance is less influenced than the bounding box based appearance under pose changes and several correctly learned parts are good enough to locate the target, our tracker still works very well in these sequences.

Complex Background. In the *football* and *stone* sequences, many similar objects confuse the trackers a lot. As shown in Figs. 2 and 3, TLD and HABT frequently skip to other similar objects. The similar appearance between the target and the background in the sequences *football*, *stone*, *tiger1* and *tiger2* makes it hard to precisely track the target. Through combining the target holistic appearance, the

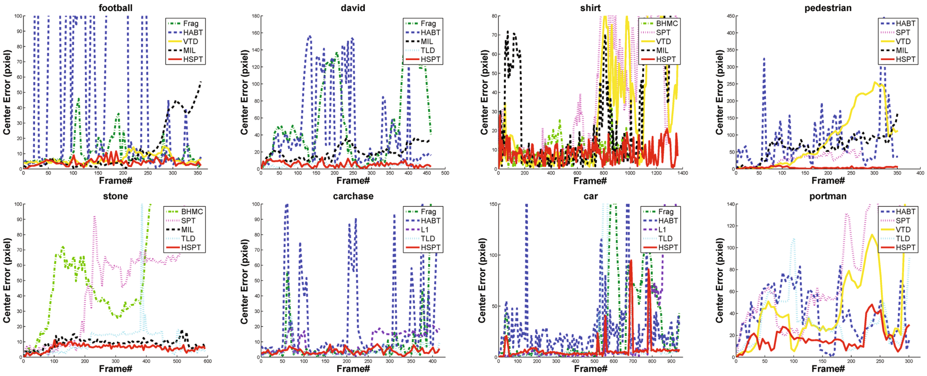

Fig. 2. Tracking results of MIL [2], VTD [7], ℓ1 [4], TLD [5], Frag [9], HABT [10], BHMC [11], SPT [14] and the proposed HSPT tracker. The results of five trackers with relatively better performance are displayed.

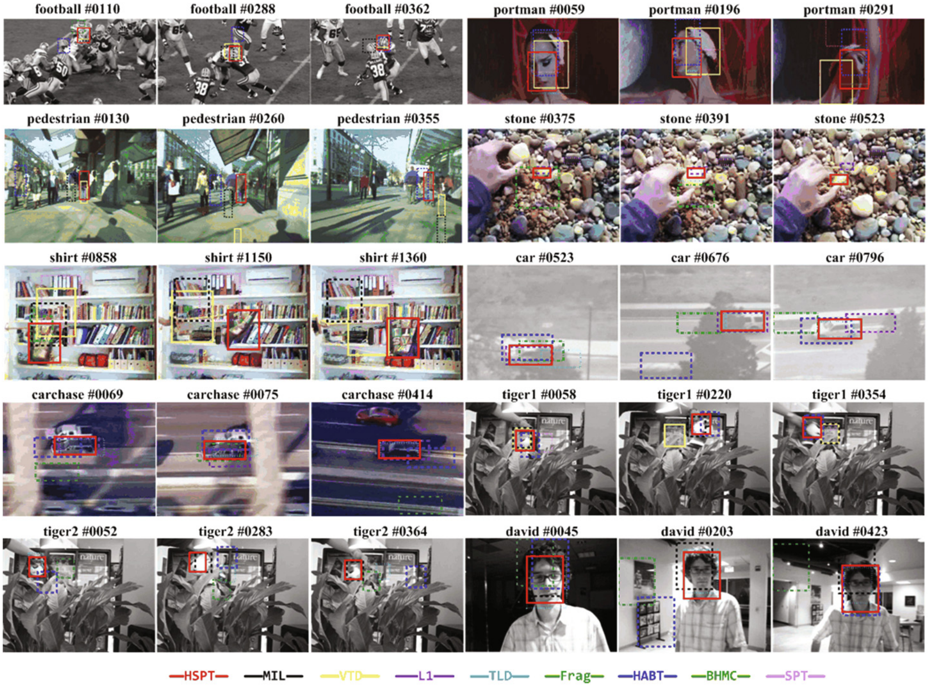

Fig. 3. Tracking results of different trackers. Only the trackers with relatively better performance are displayed.

detailed appearance of parts and the structure information between different parts, our tracker can discriminate the target from the complex background. While some specific trackers outperform ours in some specific sequences, but comprehensively speaking, our tracker performs the best.

6 Conclusion

In this paper, a novel online learned discriminative part-based tracker is proposed. The appearance of the target is described by the combination of multiple learned discriminative structural parts. In the parts learning phase, we utilize the MH algorithm based optimization framework integrated with the online SVM to infer the optimal parts state. We introduce the dynamically constructed pairwise MRF to model the interaction between neighboring parts. The experiments demonstrate the superiority of the proposed method. In the future, we will optimize the codes to make the tracker run in real-time.

Acknowledgement. This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02 B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

1. Lim, J., Ross, D.A., Lin, R.S., Yang, M.H.: Incremental learning for visual tracking. In: NIPS (2004)
2. Babenko, B., Yang, M.H., Belongie, S.J.: Visual tracking with online multiple instance learning. In: CVPR, pp. 983–990 (2009)
3. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.Z.: Robust online learned spatio-temporal context model for visual tracking. IEEE Trans. Image Process. **23**, 785–796 (2014)
4. Mei, X., Ling, H.: Robust visual tracking using $\ell 1$ minimization. In: ICCV, pp. 1436–1443 (2009)
5. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: bootstrapping binary classifiers by structural constraints. In: CVPR, pp. 49–56 (2010)
6. Wen, L., Cai, Z., Yang, M., Lei, Z., Yi, D., Li, S.Z.: Online multiple instance joint model for visual tracking. In: IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 319–324 (2012)
7. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: CVPR, pp. 1269–1276 (2010)
8. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.Z.: Online spatio-temporal structural context learning for visual tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 716–729. Springer, Heidelberg (2012)
9. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR, pp. 798–805 (2006)
10. Shahed, S.M.N., Ho, J., Yang, M.H.: Visual tracking with histograms and articulating blocks. In: CVPR (2008)
11. Kwon, J., Lee, K.M.: Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In: CVPR (2009)
12. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: ICCV, pp. 81–88 (2011)

13. Cehovin, L., Kristan, M., Leonardis, A.: An adaptive coupled-layer visual model for robust visual tracking. In: ICCV, pp. 1363–1370 (2011)
14. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: ICCV, pp. 1323–1330 (2011)
15. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: CVPR, pp. 1815–1821 (2012)
16. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: CVPR, pp. 1838–1845 (2012)
17. Cai, Z., Wen, L., Yang, J., Lei, Z., Li, S.Z.: Structured visual tracking with dynamic graph. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 86–97. Springer, Heidelberg (2013)
18. Yao, R., Shi, Q., Shen, C., Zhang, Y., van den Hengel, A.: Part-based visual tracking with online latent structural learning. In: CVPR, pp. 2363–2370 (2013)
19. Hastings, W.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57**, 97–109 (1970)
20. Bordes, A., Ertekin, S., Weston, J., Bottou, L.: Fast kernel classifiers with online and active learning. J. Mach. Learn. Res. **6**, 1579–1619 (2005)
21. Khan, Z., Balch, T.R., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1805–1918 (2005)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893 (2005)
23. Yang, M., Liu, Y., Wen, L., You, Z., Li, S.Z.: A probabilistic framework for multitarget tracking with mutual occlusions, pp. 1298–1305 (2014)
24. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Ssstrunk, S.: SLIC Superpixels. Technical report (2010)
25. Oron, S., Bar-Hillel, A., Levi, D., Avidan, S.: Locally orderless tracking. In: CVPR, pp. 1940–1947 (2012)
26. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR, pp. 1822–1829 (2012)
27. Everingham, M., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)