Online Spatio-temporal Structural Context Learning for Visual Tracking

Longyin Wen, Zhaowei Cai, Zhen Lei, Dong Yi, and Stan Z. Li*

CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun Donglu Beijing 100190, China
{lywen,zwcai,zlei,dyi,szli}@cbsr.ia.ac.cn
http://www.cbsr.ia.ac.cn

Abstract. Visual tracking is a challenging problem, because the target frequently change its appearance, randomly move its location and get occluded by other objects in unconstrained environments. The state changes of the target are temporally and spatially continuous, in this paper therefore, a robust Spatio-Temporal structural context based Tracker (STT) is presented to complete the tracking task in unconstrained environments. The temporal context capture the historical appearance information of the target to prevent the tracker from drifting to the background in a long term tracking. The spatial context model integrates contributors, which are the key-points automatically discovered around the target, to build a supporting field. The supporting field provides much more information than appearance of the target itself so that the location of the target will be predicted more precisely. Extensive experiments on various challenging databases demonstrate the superiority of our proposed tracker over other state-of-the-art trackers.

Keywords: Spatio-temporal, context constraint, subspaces learning, multiple instance boosting, unconstrained environments.

1 Introduction

Visual tracking attracts lots of attentions due to its core status in applications, *e.g.* human-computer interaction, video surveillance, virtual reality, etc. For most of these applications, trackers are demanded to work for a long time in unconstrained environments, which greatly challenges the robustness of the trackers. To overcome this difficulty, numerous complex models are designed, but most of them still focus on the appearance of target itself (*e.g.* color, edge responses, texture and shape cues) [1,2] or the difference between the target and background [3,4,5,6,7].

In real-world, the temporal and spatial information is important and necessary in tracking task. In continuous frames, the target appearance changes gradually, and all of the historical appearance variations in pose, scale and illumination have more or less influences and constraints on the next appearance state. For example, no matter what appearance changes happen to a panda, it is still a panda and the tracker should not recognize it as another animal. Meanwhile, the target moves gradually from one location to another location, rather than abruptly and discretely jumps. In another words,

* Corresponding author.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part IV, LNCS 7575, pp. 716–729, 2012.

the spatial context presents strong or weak spatial correlation between the target and the background. For example, if two similar pandas walk together, it is easy to jump from one panda to another for the trackers which only focus on appearance features. However, if the spatial context constraints are considered, the skip problem will be circumvented because the surroundings around the two pandas are different. Unfortunately, the spatio-temporal context information has not been paid enough attention in the previous tracking strategies. In this paper, we propose a novel tracking framework based on the spatio-temporal structural context to precisely predict the location of the target, which is expected to be more robust than the previous methods.

1.1 Related Works

In recent decades, numerous tracking strategies have been proposed in literatures, which perform well in some specific conditions. To better represent the target features, some methods [1,2,8,9,10] model the appearance of the target in a generative way. Fragment-based tracker [2] represents the target with histograms of local patches, which takes structural information of the target itself and handles partial occlusion very well. However, its template is not updated over time and the correlation of target and surroundings is not constructed. In [1], an Incremental Visual Tracker (IVT) adaptively updates its appearance model with the historical and sequential appearance variations. While IVT performs well in deformable motion and illumination variation, the lack of spatial information results in drift problem because the accumulated errors decrease the accuracy of appearance model.

Some discriminative model [11,5,12] formulate the tracking task as a classification problem which focuses on the difference between the target and the background. However, these trackers discard the historical separating function during updating which leads the insufficient temporal information to predict next state. Yu et al. [4] combined the generative model and discriminative model to describe different views of the target. Experimentally, the combined tracker achieves more stable performances than single generative or discriminative tracker as the result of mutual supervision. Nevertheless, the tracker in [4] just incorporated the background information as negative samples for training the classifier, and no semantic context is considered. Recently, tracking-by-detection methods [3,7,6] are very popular and reliable in long term surveillance sequences, because the appearance model will be corrected by detector over time and the target will be re-located even if it has been out of view. However, these detection based trackers are easily distracted by other objects that have similar appearance with the target, which is the result of lacking strong spatio-temporal constraints.

For long-term tracking task in unconstrained environment, merely learning the descriptive or discriminative features of the target cannot ensure the robustness of the system. Yang et al. constructed a context-aware tracker (CAT) [13] to track random field around the target instead of the target itself. The introduction of auxiliary objects that are suitable for tracking and have consistent motion correlations to the target greatly prevents the tracker from being trapped into drifting problem. Amir Saffari et. al [14] proposed a novel multi-class LPBoost algorithm to handle the tracking task. They treated the tracking task as a multi-class classification problem where background patterns become virtual classes. The proposed method performs well in constrained

environments, but it fails to handle the complex environments, *e.g.* occlusion, background clutter and illumination variations. Similar as [13], in [15], Gu and Tomasi considered the spatial relation between the similar target and track these similar targets simultaneously. However, the method ignores the temporary information of the target which causes its sensitiveness to target appearance changes and it may collapse when motion blur occurs due to the utilization of SIFT descriptors. Grabner et al. [16] introduced the definition of supporters which are useful features to predict the target location. The tracker in [16] utilizes strong motion coupling constraints to locate the target even when the target is invisible, with the help of some other available related context information. However, its detecting and matching all of the local features are expensive and the motion of the object of the view is not easily predicted. To further expand the theory of supporter, Dinh et al. developed a new context framework based on distracters and supporters [17]. The distracters are the regions that have similar appearance as the target and the supporters are the local key-points around the target which have the motion correlation with the target in a short time span. Although the introduction of context in these trackers expands the available information we can get from the scene, the motion correlation between the target and the context is hard to define.

1.2 Our Approach

The novel spatio-temporal structural context based tracker (STT) we build here greatly differs from the previous published models. For temporal context part, a new incremental subspace model is constructed to represent the gist of target with low dimensionality feature vectors, in which several sequential positive samples are packed into one subspace to update the model. Most of the appearance information of the target, including pose, scale, and illumination are efficiently incorporated into the model to help predict the next state of the target, as shown in the left side of Fig. 1. For the spatial context part, we introduce the contributors that are the regions having the same size and consistent motion correlation with the target. The positions of these contributors are produced by the key-point detection method SURF [18], which represent more information than those non-key-points. Based on the success of Fragment Tracker [2], we also decompose the target and the contributors into several small blocks. In another words, the intra-structural information and the inter-structural features are incorporated. In unconstrained environment, it is not easy to dig out the strong contextual contributors to help locate the target. Instead, numerous weak contextual contributors around the target can be combined together into a strong supporting field, as shown in the right side of Fig. 1. The representative features within the strong supporting field are optimally selected by boosting method [5] from the weak features pool. The contributions and the differences of our algorithm from other previous methods are as follows:

- The global temporal context model is constructed by the linear subspace method, which is updated with continuous positive samples and the correlation between them is considered.
- The appearance information of contributors is also considered in our model, and the pairwise features are produced by the difference between target and contributors to describe the spatial correlations.

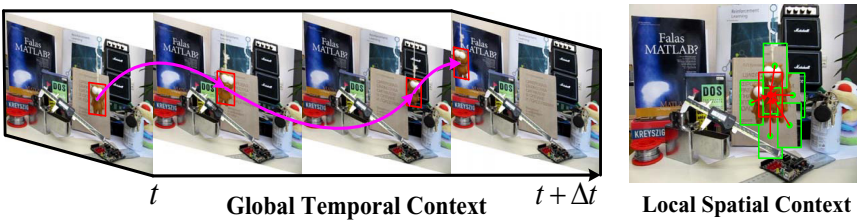

Fig. 1. The instruction of the temporal context constraint and the spatial context constraint of tracking task

- The target and contributors are decomposed into small blocks, hence the intra- and inter- structural information is described.
- Instead of building complex motion models to represent the correlation between the target and contributors, our approach efficiently utilizes boosting method to select the most representative weak relations to construct a strong supporting field.

2 MAP Spatio-temporal Structure Context Based Tracker

The tracking task is formulated as a state estimation problem and the motion process is assumed to be a Markovian state transition process. Let $O_{1:i} = \{O_1, \cdots, O_i\}$ represent the observation data set up to time i. Z_i is the state of the target at time i, which contains the position and size information of the target. In our tracker, the state vector Z_i is composed by the position of the target centered at $l_t = (x_t, y_t)$, target width w_t, target height h_t, which is defined as $Z_t = (x_t, y_t, w_t, h_t)$. The posterior probability is estimated as the recursive equation:

$$p(Z_t|O_{1:t}) \propto p(O_t|Z_t) \int p(Z_t|Z_{t-1}) p(Z_{t-1}|O_{1:t-1}) dZ_{t-1} \tag{1}$$

where $p(O_t|Z_t)$ is the likelihood of the candidate samples provided by our spatio-temporal structural context constraint. $p(Z_t|Z_{t-1})$ is the state transition probability and $p(Z_{t-1}|O_{1:t-1})$ is the state estimation probability given all observations up to time $t-1$. Similar as [5], we adopt the simplest greedy Maximum A Posteriori probability (MAP) strategy to solve the above equation, where the motion model is specified as:

$$p(l_t|l_{t-1}) = \begin{cases} 1 & \|l_t - l_{t-1}\|_2 < r \\ 0 & \|l_t - l_{t-1}\|_2 \geq r \end{cases} \tag{2}$$

where l_t is the position of the target at time t, r is the search radius. The scale of the target is similarly handled as the strategy utilized in [5].

Assume there are K contributors of the target of state s, which is represented as $f(s) = \{f_1(s), \cdots, f_K(s)\}$. The appearance model of the target in Equ. 1 is defined based on the global temporal context and the local spatial context:

$$Z_t^* = \arg\max_{Z_t} p(O_t|Z_t) = \arg\max_{Z_t} \{e^{-(1-\alpha)U(Z_t) - \alpha U(Z_t|f(Z_t))}\} \tag{3}$$

where Z_t^* is the optimal state at time t, $\alpha \in (0,1)$ is the coherence parameter to balance the global temporal context constraint and the local spatial context constraint. The energy function mentioned above consists of two terms: the global temporal context constraint energy function $U(Z_t)$ and local spatial context constraint energy function $U(Z_t|f(Z_t))$. In order to avoid the unreliable updating, we set the predefined thresholds θ_s and θ_t to decide whether the spatial and temporal context models will be updated. The algorithm of the proposed tracker is summarized in Algorithm 1 and the temporal and spatial context models are detailed in the following sections.

Algorithm 1. Spatio-Temporal Structural Context based Tracker

```
Initialize target T, extract the contributors f(·).
Initialize the global temporal context model M_t and the local spatial context model M_s.
while run do
    Sample the image to get the Candidates.
    for all Candidates do
        Calculate the global temporal context constraint energy U(Z_t);
        Calculate the local spatial context constraint energy U(Z_t|f(Z_t));
        Combine them to get the energy of the Candidates (Eq. 3)
    end for
    Find the MAP solution of the Candidates to get the minimum energy state Z_t^* (Eq. 3).
    if U(Z_t^*|f(Z_t^*)) < θ_s and U(Z_t^*) < θ_t then
        Update contributors around of the target state Z_t^*.
        Update the global temporal context model M_t with the optimal target state Z_t^*.
        Update the local spatial context model M_s with the generated contributors.
    end if
end while
```

3 Global Temporal Context with Incremental Subspace Model

Target tracking is a physically and psychologically continuous process, hence all of the prior information will be used to predict the next state of the target. The following appearances of the target have more or less correlation to the previous appearance information. For example, a man cannot abruptly change into a monkey based on historical appearances. Under this premise, global temporal context exploits historical appearance variations as an extra source of global constraints to estimate the configuration of the target. Murphy et al. [19] exploit context features using a scene 'gist', which influences priors of the object existence and state, and the work of Torralba et al. [20] shows 'gist' is sufficient to provide a useful prior for what types of objects may appear in the image. This opens our mind that we also can use object 'gist' to constrain the following states of the target. Here, we define the 'gist' as the feature vector that summarizes the target. A newly proposed incremental linear subspace method is used to reduce the high dimensionality of the feature space, so that more historical information will be stored and used efficiently.

Unlike the Hall's subspace learning method [21] and its variant [1], the newly proposed subspace learning strategy updates the energy dissipation of subspace dimension reduction in the updating process (Algorithm 2), which acquires the target features more accurately. Meanwhile, it utilizes the combined samples in adjacent frames rather than individual ones for updating. The proposed method is called Incremental Multiple Instance Subspace Learning (IMISL), which can eliminate the homogeneous noise in sequential samples effectively. An observed instance $O_t \in \mathbb{R}^d$ is a vectorized image patch corresponding to the state Z_t and d is the feature dimension of the observations. Let $\Omega_t = (\mu_t, V_t, \Lambda_t, n_t)$, where μ_t, V_t, Λ_t and n_t represent the mean vector, the eigenvectors, the eigenvalues and the number of samples of the subspace at time t respectively. Let $\Lambda_t = (\lambda_{1,t}, \cdots, \lambda_{q,t})$. To evaluate the probability of a candidate belonging to the subspace, similar to [22], the following equation is utilized:

$$U(Z_t) = \frac{\varepsilon(O_t)^2}{2\sigma_t^2} + (d-q)\log\sigma_t + \sum_{i=1}^{q}\left(\frac{G_{i,t}^2}{2\lambda_{i,t}} + \frac{1}{2}\log\lambda_{i,t}\right) \tag{4}$$

where q is the reduction dimension of the subspace, $\varepsilon(O_t) = \|O_t - VV^T O_t\|_2$ is the projection error of the candidate sample, σ_t is the energy dissipation in dimension reduction of covariance matrix at time t and $G_t = (G_{1,t}, \cdots, G_{q,t}) = V_t^T(O_t - \mu_t)$.

The core problem in incremental subspace learning is the updating strategy. Our proposed strategy utilizes the subspaces for updating instead of single samples, namely merges the two subspaces into one subspace. We first compress D updating instances into a local subspace. The subspace construction process can be completed by Eigenvalue Decomposition (EVD) or the efficient Expectation Maximization (EM) algorithm proposed in [23]. A η-truncation is utilized to decide the reduction dimension of the subspace to maintain the energy, that is $q = \arg\min_i(\frac{\sum_i \lambda_i}{tr(\Lambda)} \geq \eta)$. We derive from the basic equations of the mean value and covariance matrix of the training data, that are: $\mu^{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathcal{I}_i$, $S^{(k)} = \frac{1}{k}\sum_{i=1}^{k}(\mathcal{I}_i - \mu^{(k)})(\mathcal{I}_i - \mu^{(k)})^T$, where $S^{(k)}$ represents the covariance matrix of the subspace, $\mathcal{I}_i$ is the updating sample and $\mu^{(k)}$ is the mean value of the samples. We get the covariance matrix of the merged subspace:

$$S^{(k+l)} = \frac{k}{k+l}S^{(k)} + \frac{l}{k+l}S^{(l)} + yy^T \tag{5}$$

where $y = \sqrt{\frac{k\cdot l}{(k+l)^2}}(\mu^{(k)} - \mu^{(l)})$. Furthermore, the covariance matrix can be decomposed as the following: $S^{(k)} = \sigma_k^2 I + \sum_{i=1}^{q_k}(\lambda_{i,k} - \sigma_k^2)v_{i,k}v_{i,k}^T$, where $\sigma_k^2 = \frac{1}{d_k - q_k}\sum_{q_k+1}^{d_k}\lambda_{i,k}$, and q_k is the reduction dimension. Then plug the equation to (5):

$$S^{(k+l)} = \frac{k\sigma_k^2 + l\sigma_l^2}{k+l}I + \frac{k}{k+l}\sum_{i=1}^{q_k}(\lambda_{i,k} - \sigma_k^2)v_{i,k}v_{i,k}^T + \frac{l}{k+l}\sum_{i=1}^{q_l}(\lambda_{i,l} - \sigma_l^2)v_{i,l}v_{i,l}^T + yy^T \tag{6}$$

where $v_{i,k}$, $\lambda_{i,k}$, σ_k and $v_{i,l}$ $\lambda_{i,l}$, σ_l are the i^{th} eigenvector, i^{th} eigenvalue, energy dissipation in dimension reduction of the covariance matrix $S^{(k)}$ and $S^{(l)}$ respectively. We reformulate the Equ. 6 to get:

$$S^{(k+l)} = \frac{k\sigma_k^2 + l\sigma_l^2}{k+l}I + LL^T \tag{7}$$

where $L = [\sqrt{\rho(\lambda_{1,k} - \sigma_k^2)}v_{1,k}, \cdots, \sqrt{\rho(\lambda_{q_k,k} - \sigma_k^2)}v_{q_k,k}, \sqrt{(1-\rho)(\lambda_{1,l} - \sigma_l^2)}v_{1,l}, \cdots, \sqrt{(1-\rho)(\lambda_{q_l,l} - \sigma_l^2)}v_{q_l,l}, y]$ and $\rho = \frac{k}{k+l}$.

Due to the computation complexity of decomposing matrix LL^T directly, we decompose $L^T L$ instead, to get the decomposition of matrix $S^{(k+l)}$. Let $Q = L^T L$. The size of matrix Q is $q \times q$, where $q = q_k + q_l + 1$. We utilize the partitioned matrix to represent the matrix $Q = \begin{pmatrix} \Sigma & \beta \\ \beta^T & \alpha \end{pmatrix}$, where $\Sigma = \begin{pmatrix} \Sigma_1 & A \\ A^T & \Sigma_2 \end{pmatrix}$,$\alpha = y^T y$ and

$$\beta_i = \begin{cases} \sqrt{\rho(\lambda_{i,k} - \sigma_k^2)}v_{i,k}^T y & 1 \le i \le q_k \\ \sqrt{(1-\rho)(\lambda_{i,l} - \sigma_l^2)}v_{i,l}^T y & q_k < i < q \end{cases}$$

$$A(i,j) = \sqrt{\rho(1-\rho)(\lambda_{i,k} - \sigma_k^2)(\lambda_{j,l} - \sigma_l^2)}v_{i,k}^T v_{j,l}$$

$$\Sigma_1 = diag\{\rho(\lambda_{1,k} - \sigma_k^2), \cdots, \rho(\lambda_{q_k,k} - \sigma_k^2)\}$$

$$\Sigma_2 = diag\{(1-\rho)(\lambda_{1,l} - \sigma_l^2), \cdots, (1-\rho)(\lambda_{q_l,l} - \sigma_l^2)\}$$

Then the subspace updating process can be done efficiently by decomposing the matrix $L^T L$ and the process is detailed in Algorithm 2. In this way, the 'gist' features of the target can be captured efficiently and be utilized to predict the state of the target in the following frames.

Algorithm 2. The Subspace Updating Algorithm

1: Update the mean value of the subspaces, $\mu^{(k+l)} = \frac{k}{k+l}\mu^{(k)} + \frac{l}{k+l}\mu^{(l)}$.
2: Set $\rho = \frac{k}{k+l}$. Get the observation covariance matrix $S^{(k+l)} = (\rho\sigma_k^2 + (1-\rho)\sigma_l^2)I + LL^T$
3: Set $Q = L^T L = \begin{pmatrix} \Sigma & \beta \\ \beta^T & \alpha \end{pmatrix}$, the size of matrix Q is $(q+1) \times (q+1)$. Decompose Q as: $Q = U\Gamma U^T$, where $\Gamma = diag\{\xi_1, \xi_2, \cdots, \xi_{q+1}\}$, $U^T U = I$. Then $V_{q_k+q_l+1} = LU\Gamma^{-\frac{1}{2}}$, where matrix $V_{q_k+q_l+1} = [v_{1,k+l}, \cdots, v_{q_k+q_l+1,k+l}]$ is composed by the first $q_k + q_l + 1$ eigenvectors of the covariance matrix $S^{(k+l)}$.
4: The observation covariance matrix is represented as: $S^{(k+l)} = (\rho\sigma_k^2 + (1-\rho)\sigma_l^2)I + \sum_{i=1}^{q_k+q_l+1} \xi_i v_{i,k+l} v_{i,k+l}^T$. The first $q_k + q_l + 1$ eigenvalues of the covariance matrix can be updated as $\lambda_{i,k+l} = {\sigma^{(k+l)}}^2 + \xi_i$, and the sigma value is updated as $\sigma_{k+l}^2 = \frac{1}{d-q_{k+l}}(\sum_{i=q_{k+l}+1}^{q_k+q_l+1} \lambda_{i,k+l} + (d - q_k - q_l - 1){\sigma^{(k+l)}}^2)$, then ${\sigma^{(k+l)}}^2 = \rho\sigma_k^2 + (1-\rho)\sigma_l^2$, and $q_{k+l} = \arg\min_i(\frac{\sum_i \lambda_{i,k+l}}{\sum_{j=1}^{q_k+q_l+1} \xi_j} \ge \eta)$.

4 Local Spatial Context with Contributors

As discussed in Section 1.2, local spatial context information is derived from the area that surrounds the target to track (here we use surrounding patches as local context information, as shown in the left side of Fig. 1). The role of local context has been studied in psychology for the task of object detection [24,25], The study in [24] has proved the effectiveness of local context for object detection, and Sinha et al. [25] found that the inclusion of local contextual regions such as facial bounding contour substantially

improves face detection performance. Besides, the works in [13,16,17] show that the local context information including supporters and distracters will enforce the robustness of the tracker, even when the target is partially invisible. However, different from [13] which constructs complex relative motion model between the target and auxiliary objects and [17] which statistically counts the matched supporters around the target, our proposed strategy focuses on the weak correlation between every contributor and the target, and then combines them to construct a strong classifier to locate the target. Multiple instance boosting is exploited to efficiently select the most representative contributors and combines them together to build the supporting field.

For multiple instance boosting, each selected weak classifier corresponds to a weak correlation, and the correlations are combined together to vote the score (namely the spatial energy item in Equ. 3) of a candidate sample. The vote is expressed as:

$$U(Z_t|f(Z_t)) \propto -\sum_i h_t^i \tag{8}$$

where h_t^i is the i^{th} selected weak classifier at time t. Please refer to [5,26] for more details about multiple instance boosting algorithm.

Contributor Selection. For the contributor, similar to [17], we defines it as the key point around the target that can help to locate the target. Here, SURF descriptor is employed to find the contributors around the target which is generated by the fast Hessian algorithm. When updating, the SURF descriptor is generated in the rectangle around the center of the target with the width $r_d \cdot w$ and height $r_d \cdot h$, where r_d is the enlargement factor and we set $r_d \in [0.1, 0.6]$ in our experiments, w and h are the width and height of the target in the current frame respectively. If the extracted candidate contributors are more than the required ones, we randomly select some of them to be the final contributors. On the other hand, if they are inadequate, we randomly generate some more points to supplement them.

Feature Construction. In order to incorporate the structure information of the target, we try to partition the target and contributors into a few blocks, and the structure information is constructed with the relationships between each blocks. The structure information comes from two parts: one is the mutual-pairwise features between the blocks of the target and the contributors, and the other one is the self-pairwise features of inner blocks of the target itself. Then, these numerous relations are collected to build a feature pool. For simplicity, the structure features are produced by the difference between the sums of pixel values in each block. Certainly, other relation expression strategy can be considered, *e.g.*, Normalized Cross-Correlation (NCC). The structure features between the target and contributors deliver the holistic and detailed information of the supporting field.

Separately divide the target and contributors into $N = n_1 \times n_2$ blocks (we set $n_1 = 5, n_2 = 5$ in our experiments), $I(x, y)$ represents the pixel value of the image at position (x, y), and $P_i(s)$ represents the i^{th} block of the target or contributors corresponding to the target state s. Here we define the distance function $d(P_m(s_1), P_n(s_2))$ of two blocks:

$$d(P_m(s_1), P_n(s_2)) = \sum_{(i,j)\in P_m(s_1)} I(i,j) - \sum_{(i,j)\in P_n(s_2)} I(i,j) \tag{9}$$

Next, we collect all these weak relations to construct the feature pool. As defined in Section 2, the contributors of the target of the state s are $f(s) = \{f_1(s), \cdots, f_K(s)\}$. The pairwise feature pool $\mathcal{F}$ is constructed from two parts, the self-pairwise feature pool $\mathcal{F}_{sp}$ and the mutual-pairwise feature pool $\mathcal{F}_{mp}$, that is $\mathcal{F} = \mathcal{F}_{sp} \cup \mathcal{F}_{mp}$. The self-pairwise feature pool of the target itself is constructed as

$$\mathcal{F}_{sp} = \{d(P_i(s), P_j(s))|i = 1, \cdots, N; j = 1, \cdots, N; i \neq j\} \tag{10}$$

The mutual-pairwise feature pool of the target and its contributors is constructed as

$$\mathcal{F}_{mp} = \{d(P_i(s), P_j(f_k(s)))|i = 1, \cdots, N; j = 1, \cdots, N; k = 1, \cdots, K\} \tag{11}$$

Then the multiple instance boosting algorithm is utilized to select some of the most representative relations to construct the supporting field. In this paper, the weak classifier is adopted as in [11,5].

5 Experiments

5.1 Experimental Setup

We conduct some experiments to evaluate the performance of our spatial-temporal structural context based tracker. Our tracker is implemented in C++ code and runs on the standard PC platform. The tracker is evaluated on 10 publicly available sequences which contains different challenging conditions, and these sequences have been issued in previous works [5,27,7,6], which can be found in their own websites. Our tracker is initialized with the first frame and it outputs the trajectory of the target. The quantitative comparison results of IVT[1], FragTrack[2], SemiBoost[3], CoGD[4], MIL[5], PROST[6], VTD[27], TLD[7], ContextT[17] and our tracker are shown in Fig. 2, Table 1 and Table 2. More results can be found in the supplementary materials.

Parameters. The search radius r of the tracker is set in the interval $[20, 50]$. For the global temporal context model, every 5 frames are combined together to update the subspace model and the parameter $\eta = 0.99$ of η-truncation in subspace construction. For the local spatial context model, $K = 12$ contributors are generated to construct the supporting field and each of them are partitioned into 5×5 blocks. About 350 weak relations are combined together to construct the supporting field. For the positive bags, the samples are collected from the circle with the radius 8 and about 45 of the collected samples are packaged. For the negative bags, 50 samples are collected from the ring of the radius interval $[12, 40]$. The conservative updating threshold in our experiments are set as $\theta_s \in [-20, -10]$ and $\theta_t \in [10, 20]$. For the experimental results of other trackers we cite here, we utilize the default parameters which are provided in public available codes and choose the best one of 5 runs, or take the results directly from the published papers. Specifically, we reproduce the CoGD tracker in C++ code and adopt the parameters as described in [4].

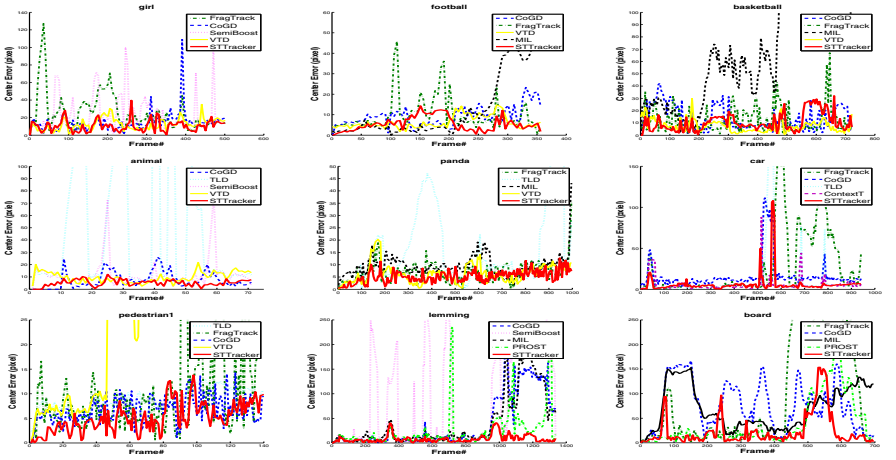

Fig. 2. Tracking results of our tracker, FragTrack[2], SemiBoost[3], CoGD[4], MIL[5], PROST[6], TLD[7], VTD[27] and Context tracker[17]. The results of five trackers with relatively better results are displayed.

5.2 Comparison with Other Trackers

Heavy Occlusion. The targets in sequence *car* and *occlude2* undergo long-term heavy occlusion for several times, and IVT which uses holistic appearances without any consideration of spatial information fails to track the target precisely. Relatively, TLD and Context Tracker perform very well in these two sequences, because the detection based trackers will re-locate the target after the occlusion, even though they lose the target during occlusion. Since the spatio-temporal context increases the possibility of our tracker to find the real target, our tracker also has good performance. A similar object usually confuses the trackers and finally misleads the trackers when it occludes the target, just like what happens in sequence *girl*. As shown in Fig. 3, approximately at the frame 463, TLD and MIL drift away for the fully occlusion of the man's face, whereas the context around the target and efficient temporal constraint provide our tracker strong discriminative ability to recognize the target.

Abrupt Motion and Motion Blur. The robustness of many trackers will be challenged by the abrupt motion resulting from hand-hold camera in sequence *pedestrian1*. The spatio-temporal context information provides enough information to ensure the robustness of the tracker. Another great challenge for the trackers is the motion blur. The loss of appearance features attributing to motion blur in the sequence *animal* and *lemming* finally results in the inaccuracy of FragTrack, SemiBoost, and TLD. However, since our temporal constraint model represents the target with low dimensionality 'gist' and the context information that can be clearly captured helps to locate the target, our tracker still has the best performance.

Cluttered Background. The cluttered background in sequence *animal* and *football* actually confuses the tracker a lot, as shown in Figure 3. Lacking spatial constraints, MIL are easily hijacked by other objects that have similar appearance with the target.

Table 1. Comparison results of average error center location in pixel

Seq.	STT	IVT	CoGD	Semi	MIL	Frag	PROST	VTD	TLD	ContextT
girl	**10.4**	40.4	14.1	22.8	31.6	25.4	19.0	12.5	35.7	18.6
occlude2	9.39	19.7	13.3	25.2	14.2	21.5	17.2	9.40	14.9	**9.25**
animal	**5.20**	226	7.38	12.3	80.3	71.4	-	9.68	50.7	81.2
basketball	**10.5**	95.4	13.8	153	93.3	12.7	-	11	158	159
football	**6.15**	17.2	9.16	102	12.7	9.92	-	6.25	13.0	51.2
pedestrian1	**5.14**	109	6.75	30.3	40.3	11.5	-	62.6	8.75	61.5
panda	**5.20**	58.2	64.5	41.7	9.42	6.85	-	6.33	17.7	77.5
car	6.26	56.9	16.6	46.4	80.7	28.6	-	51.8	11.8	**5.47**
lemming	**8.45**	128	39.8	99.8	40.5	82.8	25.1	98	167	182
board	**23.9**	169	74.5	389	69.2	90.1	39.0	70.1	134	103

Table 2. Tracking results. The numbers indicate the count of successful tracking frames based on the evaluation metric of PASCAL VOC object detection[28] in which the overlap ratio larger than 0.5 is regarded as successfully detected.

Seq.	Frames	STT	IVT	CoGD	Semi	MIL	Frag	PROST	VTD	TLD	ContextT
girl	**502**	497	353	482	388	378	378	447	**502**	219	328
occlude2	**812**	797	583	767	548	**807**	618	665	792	712	687
animal	**71**	**71**	3	62	56	5	13	-	66	43	48
basketball	**725**	**715**	75	335	90	175	630	-	601	15	50
football	**362**	346	246	292	65	272	302	-	**357**	272	55
pedestrian1	**140**	113	4	**135**	35	71	92	-	45	80	27
panda	**1000**	**580**	120	175	375	195	465	-	510	315	300
car	**945**	**915**	414	804	504	101	644	-	571	878	896
lemming	**1336**	**1246**	284	907	733	882	733	942	471	234	40
board	**698**	**583**	30	279	105	354	474	524	274	95	60

Although TLD considers positive and negative constraints and Context Tracker incorporates semantic context, they still frequently skip to other objects because they depend too much on detectors. The complex background in sequence *board* and *lemming* significantly increases the difficulty in tracking task. This is also the reason why many trackers which ignore background information including FragTrack, IVT and VTD perform bad in these sequences. Although CoGD, MIL, and PROST take the background into account, their performances are not as accurate as ours.

Large Variation of Pose and Scale. Some trackers such as FragTrack does not update their model effectively and easily lose the target when 3D pose of the target changes dramatically, as seen in sequence *girl*, *board*, and *lemming*. IVT, CoGD, and VTD adopt online updating mechanism to learn the different appearances of the target, but the large pose variation still drives them to drift away and they cannot recover. TLD and Context Tracker are good at long term surveillance sequence, but they cannot track the target precisely once large pose variation happens. When non-rigid motion happens in sequence *panda* and *basketball*, IVT and SemiBoost perform bad. Some other trackers such as CoGD, MIL and TLD have relatively good tracking results, but they do not succeed all the time. Since VTD combines multiple basic models with different features of the target, it performs well in these two sequences. Nevertheless, it does not consider the surrounding information, thus its tracking performances are not satisfactory as ours, as described in Table 1 and Table 2.

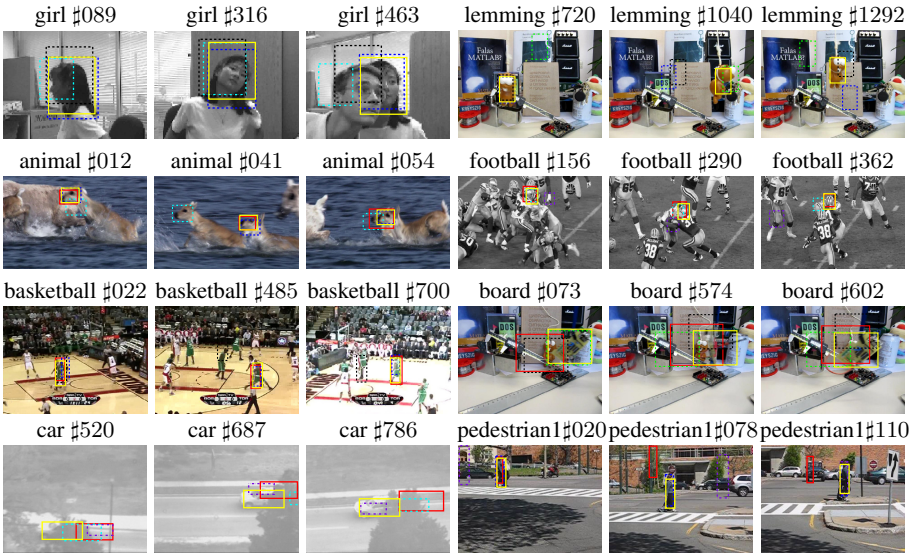

Fig. 3. Tracking results. The results of our tracker, CoGD[4], MIL[5], PROST[6], TLD[7], VTD[27] and ContextT[17] are depicted as yellow, blue, black, light green, cyan, red and purple rectangles respectively. Only the trackers with relatively better performances of each sequences are displayed.

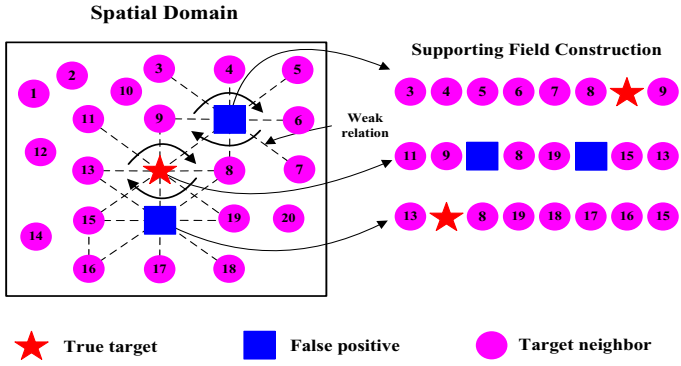

Fig. 4. The red pentagram represents the true target position, the blue triangle represents the false positive in the background and the magenta circle represents other surrounding patches. The relation between the target and its surroundings can greatly enhance the discriminability of the tracker.

5.3 Analysis

In these sequences, our proposed spatio-temporal structural context based tracker outperforms some of the state-of-the-art trackers [1,2,3,4,5,6,7,27,17]. The reason why our

STT is so stable is the introduction of global and local constraints, namely temporal and spatial context. The linear subspace (the global temporal context constraint) represents the historical appearance variations of the target with low dimensionality feature vectors. Only the gist of the object will be preserved and other noise and valueless information will be discarded during the process of subspace construction. Therefore, it is easy to explain why STT is able to handle illumination variation, motion blur, and appearance changes, because these annoying factors nearly will not influence the accuracy of our temporal context model. Particularly, we also can notice that STT is very good at dealing with the distraction by other objects which is similar to the target. As depicted in Fig. 4, when there exists a false positive near the target, while the appearances of the target and the false positive are highly similar, the surroundings of these two objects are totally varied. Once we incorporate the surrounding information around the target to build the supporting field, it is easy to differentiate the target from the false positive. Someone may doubt that STT will be drifted away by the surroundings if it keeps being updated with the surrounding information. Unlike TLD, Semiboot, and Context Tracker which utilize detectors to correct their trackers, STT is supervised by the temporal context which only focuses the target itself. The mutual supervision of spatio-temporal context ensures the long term stability of our STT.

6 Conclusion

In this paper, a spatio-temporal structural context based tracker is proposed. The appearance of target is described by the global temporal context information and the local spatial context information. The structured spatial context model automatically discovers the contributors around the target, and incorporates them to build a supporting field. In order to prevent our tracker from being drifted away by the surroundings, a strong temporal constraint model is included, which represents the target with low dimensionality feature vectors. Experimental comparison with the state-of-the-art tracking strategies demonstrates the superiority of our proposed tracker. Our future work includes the introduction of the adaptive balance coefficient between the global temporal context constraint and the local spatial context constraint, which will provide more robustness.

Acknowledgments. This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, National IoT R&D Project #2150510, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA http://www.tabularasa-euproject.org), 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. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR, pp. 798–805 (2006)

3. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
4. Yu, Q., Dinh, T.B., Medioni, G.G.: Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 678–691. Springer, Heidelberg (2008)
5. Babenko, B., Yang, M.H., Belongie, S.J.: Visual tracking with online multiple instance learning. In: CVPR, pp. 983–990 (2009)
6. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: PROST: Parallel robust online simple tracking. In: CVPR, pp. 723–730 (2010)
7. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: Bootstrapping binary classifiers by structural constraints. In: CVPR, pp. 49–56 (2010)
8. Mei, X., Zhou, S.K., Porikli, F.: Probabilistic visual tracking via robust template matching and incremental subspace update. In: ICME, pp. 1818–1821 (2007)
9. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision 77, 125–141 (2008)
10. Liu, B., Huang, J., Yang, L., Kulikowski, C.A.: Robust tracking using local sparse appearance model and k-selection. In: CVPR, pp. 1313–1320 (2011)
11. Grabner, H., Bischof, H.: On-line boosting and vision. In: CVPR, pp. 260–267 (2006)
12. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: ICCV, pp. 1323–1330 (2011)
13. Yang, M., Wu, Y., Hua, G.: Context-aware visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. 31, 1195–1209 (2009)
14. Saffari, A., Godec, M., Pock, T., Leistner, C., Bischof, H.: Online multi-class lpboost. In: CVPR, pp. 3570–3577 (2010)
15. Gu, S., Tomasi, C.: Branch and track. In: CVPR, pp. 1169–1174 (2011)
16. Grabner, H., Matas, J., Van Gool, L.J., Cattin, P.C.: Tracking the invisible: Learning where the object might be. In: CVPR, pp. 1285–1292 (2010)
17. Dinh, T.B., Vo, N., Medioni, G.G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: CVPR, pp. 1177–1184 (2011)
18. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.J.: SURF: Speeded-up robust features. Computer Vision and Image Understanding 110, 346–359 (2008)
19. Murphy, K., Torralba, A., Freeman, W.T.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. In: NIPS (2003)
20. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision 53, 169–191 (2003)
21. Hall, P.M., Marshall, A.D., Martin, R.R.: Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. Image Vision Comput. 20, 1009–1016 (2002)
22. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. IEEE Trans. Pattern Anal. Mach. Intell. 19, 696–710 (1997)
23. Tipping, M., Bishop, C.: Probabilistic principal component analysis. J. Royal Statistical Soc. Series B 61, 611–622 (1999)
24. Palmer, S.: The effects of contextual scenes on the identification of objects. Memory & Cognition 3, 519–526 (1975)
25. Torralba, A., Sinha, P.: Detecting faces in impoverished images. Journal of Vision 2 (2002)
26. Viola, P.A., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS (2005)
27. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: CVPR, pp. 1269–1276 (2010)
28. Everingham, M., Van Gool, L.J., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88, 303–338 (2010)