Robust Maximum Margin Correlation Tracking

Han Wang, Yancheng Bai, Ming Tang
{han.wang, ycbai, tangm}@ nlpr.ia.ac.cn

National Lab of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China

Abstract. Recent decade has seen great interest in the use of discriminative classifiers for tracking. Most trackers, however, focus on correct classification between the target and background. Though it achieves good generalization performance, the highest score of the classifier may not correspond to the correct location of the object. And this will produce localization error. In this paper, we propose an online Maximum Margin Correlation Tracker (MMCT) which combines the design principle of Support Vector Machine (SVM) and the adaptive Correlation Filter (CF). In principle, bipartite classifier SVM is designed to offer good generalization, rather than accurate localization. In contrast, CF can provide accurate target location, but it is not explicitly designed to offer good generalization. Through incorporating SVM with CF, MMCT demonstrates good generalization as well as accurate localization. And because the appearance can be learned in Fourier domain, the computational burden is reduced significantly. Extensive experiments on public benchmark sequences have proven the superior performance of MMCT over many state-of-the-art tracking algorithms.

1 Introduction

Visual tracking is a significant problem in computer vision and it has been used in various applications such as automatic object identification, automated surveillance, vehicle navigation *et al.* Visual tracking has made great progress in the last decades and there are many different tracking approaches, such as kernel based tracking [1], particle filter based tracking [2], and tracking by detection [3]. However, designing a robust tracker is still a challenging problem, as the tracking results can be greatly influenced by moving out of plane, illumination changes, occlusion [4] *et al.*

Recently, tracking by detection has become a hot topic in single object tracking [3]. It stems directly from the offline training object detection methods, and it turned the offline training to online training to solve tracking problems.

Avidan [3] uses SVM to build a classifier separating the object from the background. The classifier uses offline training SVM integrated with optical flow algorithm to locate the object. But as the classifier is offline trained, the tracker can not adapt to the appearance changes of the object. In order to solve this problem, ensemble tracking [5] algorithm has been proposed. The algorithm collected positive and negative samples from the object and background regions to

train the weak classifiers, and used adaboost to select the most effective weak classifiers. A weighted sum of the selected classifiers presents the final strong tracker. As selecting appropriate positive and negative samples can influence the tracking results a lot, Babenko [6] propose a more robust algorithm based on multiple instance learning. The algorithm is more robust and have more fault tolerance as instead of receiving a set of instances which are labeled positive or negative, the learner receives a set of bags that are labeled positive or negative. As wrong labeling can be always occurred in tracking, Z.Lalal proposed [7] using 'P-N learning' to estimate the samples that are wrong labeled. The tracker utilizes P-*expert* to find the wrong labeled positive samples and N-*expert* to find the wrong labeled negative samples.

All of the aforementioned algorithms have one thing in common, in training process, they all regarded the tracking problem as a bipartite classification problem. This can severely influence the localization performance of the tracker. Assume in frame t, if we have a d-dimensional solution vector $\mathbf{w}$, correlating it with the image search patch, the peak of the response map can represent the object center. The ideal response map obtains a sharp correlation peak, which is centered at the object center. However, the response map of these trackers usually exhibits very broad peaks as they use binary labels for training. Broad peak will cause poor localization performance, as the top of the peak may be spread over several pixels thus can not correspond to the target center. Hare proposed structured SVM tracker[8] labeling every sample differently to improve the localization performance. The training process of all aforementioned trackers is calculated in spatical domain. Thus these trackers can not choose dense sampling strategy which will becomes computational burdens. Instead, they choose sparse sampling as shown in Fig.1(a): positive samples are usually randomly collected in the target's neighbour, which can make the results severely influenced by the selection of samples.

Bolme [9] proposed the MOSSE tracker using the adaptive Correlation Filter for tracking. It uses dense sampling strategy, shown in Fig.1(b). And as the center patch labels 1 and the value of labels degrades as the distance between the sample and the target center increases. This strategy can keep the structure of the target and localize accurately. As the model is computed in its Fourier domain, the computational burden can be reduced a lot. CF can generate sharp peaks and thus provide good localization performance, but they are not explicitly designed to offer good generalization.

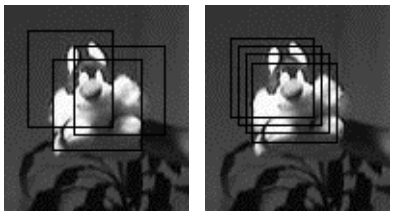
While SVM are designed to maximize the margin of different classes, it usually has good generalization performance. In priciple, combining the design of CF and SVM, Andres[10] proposed the MMCF, an offline training algorithm for object detection. The classifier has good generalization and localization performance rather than SVM and CF. And it can be processed in Fourier domain for fast training. But the MMCF is an offline training method, in tracking, it can not adapt to the appearance changes of the target.

In this paper, however, we propose the MMCT. This tracker integrates the design of CF and SVM, using the two criteria to build the objective function.

The tracker uses dense sampling strategy around the target, and away from the target, it randomly sample the negative patches. By using the SVM to constrain the coefficients of CF, the response map of the tracker can produce discriminative sharp peaks around the target center and small values away from the target center.

Different from MMCF, we achieve an online learning algorithm. Instead of using a weighted sum of models to update the tracker as many traditional tracker do, we import the previous model into the objective function. Through incorporating the last model into the SVM constraint, the tracker of consecutive frames can maintain continuity and at the same time achieves good generalization. This makes the tracking model more robust, and as the objective function can be processed into Fourier domain, the computational burden can be dramatically reduced.

The rest of the paper is organized as follows. In Sec 2, we give a brief introduction for the MMCF and then present our algorithm in details. Experiments and the results of comparing with other state-of-art algorithms are shown in Sec 3. In Sec 4, we will summarize our work.



(a) (b)

Fig. 1. This figure illustrates the sparse sampling strategy and dense sampling strategy. (a) illustrates the sparse sampling, it randomly samples p windows and save them; (b) illustrates the dense sampling, it samples all subwindows together and save one image.

1.1 Tracking model

In this section, we will introduce our online tracking algorithm. In Sec 2.1, we give a brief introduction of how CF works in tracking, and in Sec 2.2, the offline training model used in object detection is represented. In Sec 2.3, we introduce our online updating model and in Sec 2.4, a detailed tracking strategy will be given.

1.2 The adaptive correlation filter

As mentioned above, many traditional trackers use sparse sampling strategy, it means that several positive patches are randomly sampled around the object

and all labeled 1. Obviously, there is a lot of redundancy because of the overlap between samples. Besides, as the labels of positive samples are all ones, it ignores the structure of the target, which can cause poor localization performance.

The Adaptive Correlation Filter [11] is firstly rooted on classical signal processing, and now widely used in localization and classification. It realizes dense sampling strategy around the object and at the same time, as it labels each sample differently, the model can present the structure of the target.

We start a general formulation to introduce the notation. First, we introduce the notation of circulant matrix [12]. If a matrix is circulant, means that if a n*n matrix $C(\mathbf{u})$ is extracted from the $n * 1$ vector $\mathbf{u}$ by concatenating all possible cyclic shifts of $\mathbf{u}$,

$$C(\mathbf{u}) = \begin{pmatrix} u_0 & u_1 & u_2 \cdots & u_{n-1} \\ u_{n-1} & u_0 & u_1 \cdots & u_{n-2} \\ u_{n-2} & u_{n-1} & u_0 \cdots & u_{n-3} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ u_1 & u_2 & u_3 \cdots & u_0 \end{pmatrix}. \tag{1}$$

Since the product $C(\mathbf{u})\mathbf{v}$ can be seen as the convolution of the two vectors $\mathbf{u}$,$\mathbf{v}$,we can compute it in Fourier domain, using

$$\widehat{C(\mathbf{u})\mathbf{v}} = \hat{\mathbf{u}}^* \odot \hat{\mathbf{v}} \tag{2}$$

where $\odot$ denotes the element-wise product, and $\hat{}$ denote Fourier transform, and * represents the complex-conjugate.

The dense sampling strategy at many subwindows in our paper is conceptually close to circulant matrix. In frame t, there are N target image patches from the last N frames $P_{t-N+1}, P_{t-1}, ..., P_{t-1}, P_t \in \mathbf{R}^{m*k}$. For each patch P_i, The dense sampling subwindows and their labels are $(\mathbf{x}_{i1}, y_{i1}), (\mathbf{x}_{i2}, y_{i2}), (\mathbf{x}_{ij}, y_{ij}) \ldots (\mathbf{x}_{id}, y_{id})$, $d = m * k$, where $\mathbf{x}_{ij}$ can be seen as a shifted vectorized version of image patch P_i, while y_{ij} means the label of $\mathbf{x}_{ij}$. As a linear classifier can be seen as $f(x) = \mathbf{w}^T * \mathbf{x} + b$, ignore the bias term b , just as [12] do, with quadratic loss, the objective minimization problem can be simply seen as

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \|\mathbf{w}^T B_i - \mathbf{g}_i\|^2 \tag{3}$$

where $B_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{id}]$; $\mathbf{g}_i = [y_{i1}, y_{i2}, ..., y_{id}]^T$. Unlike traditional labeling strategy, in order to output sharp peaks, instead of using binary labels, the model uses a Gaussian function-like to represent $\mathbf{g}_i$ whose peak is at the object center. As the structure of B_i , the sampling subwindows $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{id}$, is close to circulant matrix. So the Fourier transform of the Eq.3 is as follows,

$$\min_{\hat{\mathbf{w}}} \sum_{i=1}^{N} \|\hat{\mathbf{w}} \odot \hat{\mathbf{x}}_i - \hat{\mathbf{g}}_i\|^2 \tag{4}$$

where $\hat{\mathbf{x}}_i$ is the vectorized version of 2-D Fourier transform of the image patch P_i . In tracking process, when the tracker $\mathbf{w}$ is correlated with the test image, the ideal response map $\mathbf{g}_t$ can obviously produce sharp peak to localize correctly. But as the tracker is not designed for classification, when the background clutters, it may not track well.

1.3 Offline training model

Trackers related to adaptive Correlation Filter are MOSSE [9] and Circulant [12] trackers, they have fast speed in tracking.

While in object detection, the SVM classifier is designed to maximize the margin and can always produce robust classifiers to classify the positive and negative samples. As the training samples are binary labeled, the output, which is resulting from cross-correlation of SVM templates with testing images, can not produce sharp peaks. As mentioned above, this will cause poor localization performance. Andres[10] propose an offline training object detection algorithm, MMCF. The MMCF uses two criteria combining the design of the SVM and CF. We first follow the notation in [10] to introduce the model.

The MMCF classifier is a multi-criteria classifier. The first criterion is SVM. Given N of training column vectors $\mathbf{x}_i \in \mathbb{R}^d$ and the class labels $t_i \in \{-1, 1\}$ $\forall i \in 1, \ldots, N$, the objective function of SVM can be expressed as follows,

$$\min_{\mathbf{w},b} \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i$$

$$s.t.\ t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq c_i - \xi_i \tag{5}$$

The second criterion is the CF, just as mentioned above, the objective function is $\min_{\mathbf{w}} \sum_{i=1}^{N} \|\mathbf{w}^T B_i - \mathbf{g}_i\|^2$, where $\mathbf{g}_i = [0, \ldots, 0, \mathbf{w}^T\mathbf{x}_i, 0, \ldots, 0]$, we prefer the center of the object is $\mathbf{w}^T\mathbf{x}_i$, while others close to 0. Combined with the SVM, the objective function can be seen as follows,

$$\min_{\mathbf{w},b}(\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i, \sum_{i=1}^{N} \|\mathbf{w}^T B_i - \mathbf{g}_i\|^2)$$

$$s.t.\ t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq c_i - \xi_i \tag{6}$$

where c_i=1 for positive image patches and $c_i = \varepsilon$ for negative image patches, where ε is a small value constant. That means for positive image patchs, we expect a value above 1, while for negative patches, the expected value is close to 0. The large margin of SVM means good generalization performance, while the CF criterion makes sharper correlation peak. The objective function suggests a correlation response map, which has a sharp peak at the target center and small values everywhere else.

1.4 Online tracking model and optimization

In tracking problems, we'd like to have dense sampling in the target center's neighbourhood to ensure good localization performance, at the same time, target should be separated from the background. We also need an online training strategy to adapt to the appearance changes of the object. Above these, we propose an online tracking model which can produce discriminative sharp peaks.

In our approach, suppose in frame t, after locating the object in $\mathbf{p}_t$, we extract the positive image patch P_t centered at $\mathbf{p}_t$ which has the same size with the target, and with the last k frames's k positive image patches, we have a positive training set $(P_t, P_{t-1}, \ldots, P_{t-k})$. To get the negative training set, we simply collected m patches away from the target in frame t, $(P_1, P_2, \ldots, P_m)$. We then train the online model w_{t+1} using the sample sets. Instead of using the simply weighted sum $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta\mathbf{w}$ to update the model, where $\mathbf{w}$ is the trained model using current sample sets, we optimize $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$ in SVM criterion instead of $\|\mathbf{w}\|^2$ to keep the continuity between frames. Given $N = k + 2 + m$ of training column vectors $\mathbf{x}_i \in \mathbb{R}^d$ which is the vectorized version of P_i, and the class labels $t_i \in \{-1, 1\}$ $\forall i \in 1, \ldots, N$, the online tracking model can be expressed as follows,

$$\min_{\mathbf{w}_{t+1}, b}(\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + C\sum_{i=1}^{N}\xi_i, \sum_{i=1}^{N}\|\mathbf{w}^T B_i - \mathbf{g}_i\|^2)$$

$$s.t.\ t_i(\mathbf{w}_{t+1}^T\mathbf{x}_i + b) \geq c_i - \xi_i \tag{7}$$

where $\mathbf{g}_i = [0, \ldots, 0, \mathbf{w}_{t+1}^T\mathbf{x}_i, 0, \ldots, 0]$, the nonzero value $\mathbf{w}_{t+1}^T\mathbf{x}_i$ is at the target center, and the other elements are all zeros. Just the same as the offline model, B_i represents the circulant matrix of $\mathbf{x}_i$, $c_i = 1$ for positive training set and $c_i = \varepsilon$ for negative training set. The objective function shows that in target center, we prefer a value of above 1, and the value decays to small values as the distance increases. The tracker uses dense sampling strategy around the target, so it can produce sharp peaks of the correlation output, and at the same time, using maximum margin to constrain the CF, the generalization performance improves a lot. With the $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$ constraint, the trackers of consecutive frames can maintain continuity.

In order to make use of the property that cross-correlation in the spatial domain is equivalent to multiplication in frequency domain, we transform Eq.7 to its Fourier domain. We turn the SVM to the frequency domain by using the Parseval theorem. While the correlated part can be easily transformed to the Fourier domain as shown in section 2.1. Then, Eq.7 can be transformed as follows,

$$\min_{\hat{\mathbf{w}}_{t+1}, b}(\|\hat{\mathbf{w}}_{t+1} - \hat{\mathbf{w}}_t\|^2 + C\sum_{i=1}^{N}\xi_i, \sum_{i=1}^{N}\|\hat{\mathbf{w}}_{t+1}^* \odot \hat{\mathbf{x}}_i - \hat{\mathbf{g}}_i\|^2)$$

$$s.t.\ t_i(\hat{\mathbf{w}}_{t+1}^\dagger\hat{\mathbf{x}}_i + b') \geq c_i - \xi_i \tag{8}$$

Where † is the conjugate transpose. The multi-criteria function shown in Eq.8 is formulated by two quadratic function, Refregier [13] showed that this can be optimized by minimizing a weighted sum of the two criteria, so it can be expressed as,

$$\min_{\hat{\mathbf{w}}_{t+1},b} \lambda\|\hat{\mathbf{w}}_{t+1}-\hat{\mathbf{w}}_t\|^2+\lambda C\sum_{i=1}^{N}\xi_i+(1-\lambda)\sum_{i=1}^{N}\|\hat{\mathbf{w}}_{t+1}\odot\hat{\mathbf{x}}_i-\hat{\mathbf{g}}_i\|^2$$

$$s.t.\ t_i(\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{x}}_i+b')\geq c_i-\xi_i \tag{9}$$

where λ represents the trades-off parameter between the margin criterion and localization criterion. When $\lambda=1$ it equals to SVM tracker and vice versa.

For the second part, as $\mathbf{w}_{t+1}^{T}\mathbf{x}_i$ is the same as $\frac{1}{d}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{x}}_i$, using Pascal's theorem. The Fourier transform of $\mathbf{g}_i$ is as follows,

$$\hat{\mathbf{g}}_i=\mathbf{1}*(\frac{1}{d}\hat{\mathbf{x}}_i^{\dagger}\hat{\mathbf{w}}_{t+1}) \tag{10}$$

where $\mathbf{1}$ represents a column vector whose elements are all *1.* using the diagonal matrix $\hat{X}_i\mathbf{1}=\hat{\mathbf{x}}_i$, then the right part of Eq.8 can be expressed as follows,

$$\begin{aligned}\sum_{i=1}^{N}\|\hat{\mathbf{w}}_{t+1}\odot\hat{\mathbf{x}}_i-\hat{\mathbf{g}}_i\|^2&=\sum_{i=1}^{N}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{X}_i\hat{X}_i^{*}\hat{\mathbf{w}}_{t+1}-\frac{2}{d}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{X}_i\hat{\mathbf{g}}_i+\frac{1}{d^2}\hat{\mathbf{g}}_i^{\dagger}\hat{\mathbf{g}}_i\\&=\sum_{i=1}^{N}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{X}_i\hat{X}_i^{*}\hat{\mathbf{w}}_{t+1}-\frac{2}{d}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{X}_i\mathbf{1}\hat{\mathbf{x}}_i^{\dagger}\hat{\mathbf{w}}_{t+1}+\frac{1}{d^2}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{x}}_i\mathbf{1}^{\dagger}\mathbf{1}\hat{\mathbf{x}}_i^{\dagger}\hat{\mathbf{w}}_{t+1})\\&=\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{Z}\mathbf{w}_{t+1}\end{aligned} \tag{11}$$

where

$$\hat{Z}=\sum_{i=1}^{N}(\hat{X}_i\hat{X}_i^{*}-\frac{1}{d}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^{\dagger}) \tag{12}$$

Subsume Eq.11 into Eq.9, we can rewrite Eq.9 as follows,

$$min_{\hat{\mathbf{w}}_{t+1},b}\lambda\|\hat{\mathbf{w}}_{t+1}-\hat{\mathbf{w}}_t\|^2+\lambda C\sum_{i=1}^{N}\xi_i+(1-\lambda)\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{Z}\hat{\mathbf{w}}_{t+1}$$

$$s.t.\ t_i(\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{x}}_i+b')\geq c_i-\xi_i \tag{13}$$

With one quadratic term subsumed into the other quadratic term, Eq.13 can be rewritten as follows,

$$min_{\hat{\mathbf{w}}_{t+1},b}\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{S}\hat{\mathbf{w}}_{t+1}+\lambda C\sum_{i=1}^{N}\xi_i-2\lambda\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{w}}_t$$

$$s.t.\ t_i(\hat{\mathbf{w}}_{t+1}^{\dagger}\hat{\mathbf{x}}_i+b')\geq c_i-\xi_i \tag{14}$$

where $\hat{S} = \lambda I + (1-\lambda)\hat{Z}$, as $0 < \lambda < 1$, $\hat{S}$ is positive definite matrix. And we can transform the data that $\tilde{\mathbf{w}} = \hat{S}^{\frac{1}{2}}\hat{\mathbf{w}}$ and $\tilde{\mathbf{x}}_i = \hat{S}^{-\frac{1}{2}}\hat{\mathbf{x}}_i$. So we can easily compute the dual form of Eq.14,

$$\min_{\mathbf{a}} \mathbf{a}^T T\tilde{X}^\dagger \tilde{X} T\mathbf{a} + (\mathbf{c}^T - 2\tilde{X}^\dagger T\tilde{\mathbf{w}}_t)\mathbf{a}$$

$$s.t.\ \mathbf{0} \leq \mathbf{a} \leq \mathbf{1}C', \mathbf{a}^T\mathbf{t} = 0 \tag{15}$$

where $\tilde{X} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N], \mathbf{t} = [t_1, \ldots, t_N]^T, \mathbf{c} = [c_1, ..., c_N]^T, C' = \lambda C$, and T is the diagonal matrix with $\mathbf{t}$ along the diagonal. With the dual form, we can optimize $\mathbf{a}$ using Sequential minimal optimization (SMO) [14]. SMO breaks this problem into many subproblems that each problem solve for one nonoverlap pair of $\mathbf{a} = [a_1, \ldots, a_N]^T$. It recursively solves for $\mathbf{a}$ until convergence, and after solving for $\mathbf{a}$, the tracking model $\hat{w}_{t+1}$ can be computed as follows,

$$\hat{\mathbf{w}}_{t+1} = \hat{S}^{-\frac{1}{2}}\widetilde{X}\mathbf{a} \tag{16}$$

Here as $\hat{\mathrm{S}}$ is not a diagonal matrix, so when computing the inverse of the matrix, it is very computationally expensive. As the target patch dimension d is always very large, so we can approximate $\hat{\mathrm{S}}$ as,

$$\hat{S} = \lambda I + (1-\lambda)\hat{Z} = \lambda I + (1-\lambda)\sum_{i=1}^{N}(\hat{X}_i\hat{X}_i^* - \frac{1}{d}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^\dagger)$$

$$\approx \lambda I + (1-\lambda)\sum_{i=1}^{N}\hat{X}_i\hat{X}_i^* \tag{17}$$

As the objective function can be processed in Fourier domain, the computational burden can be significantly reduced. And with the larger λ , the stronger generalization performance and smaller λ can make the model output sharper peak.

1.5 Tracking process

In frame *t*, given the model $\mathbf{w}_t$ and $\mathbf{p}_{t-1}$ the center of frame *t-1*, the prediction process is to find the new target center $\mathbf{p}_t$. We cropped the search patch 1.5 times as big as the target, centered at $\mathbf{p}_{t-1}$ in frame *t*, and correlate it with w_t, get the correlation response map $\mathbf{g}_t$. The strength of the peak of $\mathbf{g}_t$ can be measured by the Peak to Sidelobe Ratio(PSR) [9]. To compute the PSR, we first divide the response map $\mathbf{g}_t$ into two portions. The peak represents the maximum value of the response map and the sidelobe is the rest of the pixels excluding an $11 * 11$ window around the peak. The PSR can be computed as $\frac{g_{max}-\mu_s}{\sigma_s}$, where g_{max} is the peak value, and μ_s, σ_s are the average and standard deviation of the sidelobe. The PSR can be used to detect the object occlusion or tracking failure. If PSR is smaller than 6 (experience in our experiment), the target is supposed to be missing, and we will search the whole image and stop updating the model. Algorithm 1 summarizes our tracking algorithm.

Algorithm 1 Maximum Margin Correlation Tracker.

Require:
The current frame image, F_t;
The current model, $\mathbf{w}_t$;
The object center of frame *t-1*, $\mathbf{p}_{t-1}$;

Ensure:
The new tracker $\mathbf{w}_{t+1}$ and the object center of frame t $\mathbf{p}_t$;

1: Cropped the current search image patch S_t centered at $\mathbf{p}_{t-1}$;
2: Getting the current response map, using 2-D cross-correlation, where $\mathbf{x}_t$ is the vectorized version of S_t: $\mathbf{g}_t = \mathbf{x}_t \otimes \mathbf{w}_t$;
3: Compute the PSR of $\mathbf{g}_t$, and get the object center p_t according to Sec 2.4;
4: Using PSR to estimate if the object is occluded, according to Sec 2.4;
5: Sample the positive image patch P_1 in frame t and add it to positive templates. Sample negative patches $P_2, \ldots, P_n$;
6: Using the positive templates, the negative patches, and $\mathbf{w}_t$ to update the model $\mathbf{w}_{t+1}$ according to Sec.2.3;
7: **return** $\mathbf{w}_{t+1}, \mathbf{p}_t$;

2 Experiments

We evaluated our tracking system on twelve challenging videos, all of the videos come from the benchmark. These videos contain many kinds of objects (car, pedestrian, human body, faces animals *et al*).
The proposed algorithm is implemented in MATLAB on a workstation with an Intel core i5 3.2GHz processor and 16G RAM. The average pruning time is 3-4 frames per second.
In all of the experiments, the parameters are all fixed. In the training stage, we sample 40 negative image patches within 30 pixels away from the bounding box of the target. And the positive patches are collected from the last 10 frames' target patches. And we choose λ =0.15 to balance the correlation results. We will make our experiments in two ways. First, we compare our algorithm with the related algorithms. And then Comparison with other state-of-the-art algorithms are made.

2.1 Pre-processing

The proposed method uses Fourier transform in training process. As Fast Fourier Transform(FFT) is periodic, it is very sensitive to the image boundary. A noisy Fourier representation can be generated if there exists big discontinuity between opposite edges of the images. The effect can be reduced by multiplying a hanning window with the image to gradually reduce the training patches to zero.

2.2 Comparison with MOSSE and SVM

To demonstrate the improvements of our approach in localization and generalization , we first make a experiment comparing our algorithm with the SVM, and also the MOSSE tracker.

For generality, there are many kinds of objects (human body, face, rigid object and toy). And the mean center position error per frame is used as criterion. Table. 1 shows the quantitative performance of these algorithms.

It can be seen that the proposed method outperforms other trackers. Fig.3(a-c) shows the results of some typical videos under difficult situations. In the experiment,we also find the Moose filter is sensitive to the initialization bounding box, if the initialization bounding box included much background information, the tracking result can be severally influenced by background, And in SVM tracker, tracking results can be improved by increasing positive and negative samples, but this can increase the computational burden. It can be seen in Fig.3(a) that under pose changes, MOSSE and our tracker can localize the object correctly, but SVM tracker drifts. Fig.3(b-c) shows with scale changes and out of plane rotation changes, our tracker performs well compared with others.

2.3 Comparison with other trackers

In this section, we compare the proposed tracker with other 5 state-of-the-art trackers (the tracking results of them are provided by the benchmark), including the TLD [7], Struck [8], MIL [6], L1APG [15], MTT [16] trackers.

2.4 Quantitative Evaluation

We evaluate the performance of these trackers using the center location error. Table. 2 reports the average center location errors in pixels. It can be seen that under different situations, our tracker can locate accurately, it always performs best or second best. Fig.2 shows the tracking results of different trackers.

2.5 Qualitative Evaluation

Illumination,pose and Scale changes we evaluate sequences with different kinds of illumination changes. The *david* and *Trellis* contain gradual illumination, pose and scale changes. We can see from Fig.2 that under illumination changes (e.g. Trellis #51, #228)only our tracker and Struck tracker can locate the object accurately, other trackers have drifts to some extent. And when the pose changes a lot (e.g. *Trellis* #356), only our tracker performs well. In the sequence *david*, when the scale changes a lot (e.g. #482, #525), only the proposed algorithm is able to track the object accurately. This can be attributed to that we design a shape peak for the center of the target. So even the object's appearance changes, we not only can classify the object, but also find its accurate center.

Occlusion The target objects are partially occluded in the *Women, Occluded face 2, SUV* sequences. When the target is severely occluded (e.g. *SUV* #526, *Woman* #133), our tracker can still perform well. By using dense sampling strategy around the target, the spatial information is maintained a lot and thus can handle occlusion well.

Out of plane rotation and abrupt motion The target objects *Sylvster* and *football1* undergo out of plane rotation and abrupt motion. For out of plane rotation (e.g. *Sylvster*#0619, #1041), Most trackers except the proposed tracker and the **Struck** method drift. For abrupt motion and rotation out of plane (e.g. *Football1* #0038), our tracker and TLD perform well.
Background clutters In the *football1* and *cardark* sequences, the target object undergos fast movements in cluttered backgrounds. Our tracker performs well where as others fail to locate the target object.

Sequences	bolt	cardark	Suv	football1	freeman3	sylvster	Trellis	Woman	david	deer	dog1	faceocc2
MOSSE	30	4.3	42	22.6	15	10.8	11.9	16.6	12	7.5	10.3	14.6
SVM	200	6.9	50	49	50	31.7	13.3	71.4	53.5	23	7.2	14.2
Ours	**24**	**2.2**	**4.9**	**14**	**12**	**8**	**6.4**	**10**	**10.1**	**5.8**	**4.8**	**10**

Table 1. The average center location error of twelve sequences is the distances between the tracking results center and the ground truthes of them. The bold represents for the best tracker.

Sequences	bolt	cardark	SUV	football1	freeman3	Sylvster	Trellis	Woman	david	deer	dog1	faceocc2
TLD	*231*	35	56	9.7	14	77	27	*11*	34	7	22	*18.7*
Struck	250	*3.9*	41	**13**	12	*26*	*6.4*	12.2	*15*	*7*	*11*	58
MIL	286	48	*12*	32	19	90	135	27	73	9	21	83
L1APG	283	25.2	15	22	30	40	165	71	345	11	14	101.5
MTT	278	20.7	17	18	*12*	58	170	25	350	9	16	119
Ours	**24**	**2.2**	**4.9**	*14*	**12**	**8**	**6.4**	**10**	**10.1**	**5.8**	**4.8**	**10**

Table 2. Compared average center error(pixels)on twelve sequences. The **bold** represents for the best tracker, and *italic* for the second best.

3 Conclusion

In this work, we present a new adaptive tracking-by-detection method based on adaptive correlation filter and the SVM. Unlike existing method using sparse sampling strategy and focusing on classification, thinking from the intension of tracking, localizing the target, we build the model getting good performance in localization and also separate the target from the clutter background. And we transform it to Fourier domain for fast training using FFT. And in training, we do not simply use the weighted sum of the model, which are computed from different frames, representing the new model. But we change the SVM objective criteria to both adapt to new samples and also keep consistence with previous model. Through experiments on public benchmark sequences, we also clearly demonstrated that our algorithm can track objects very well under large pose, scale variation ,occlusion and cluttered background. And our MMCT can almost always outperform the state-of-the-art algorithms.

Fig. 2. Tracking results on six of the twelve vedios(*faceocc2, SUV, cardark, david, sylvster, football1, Threllis, Woman*).

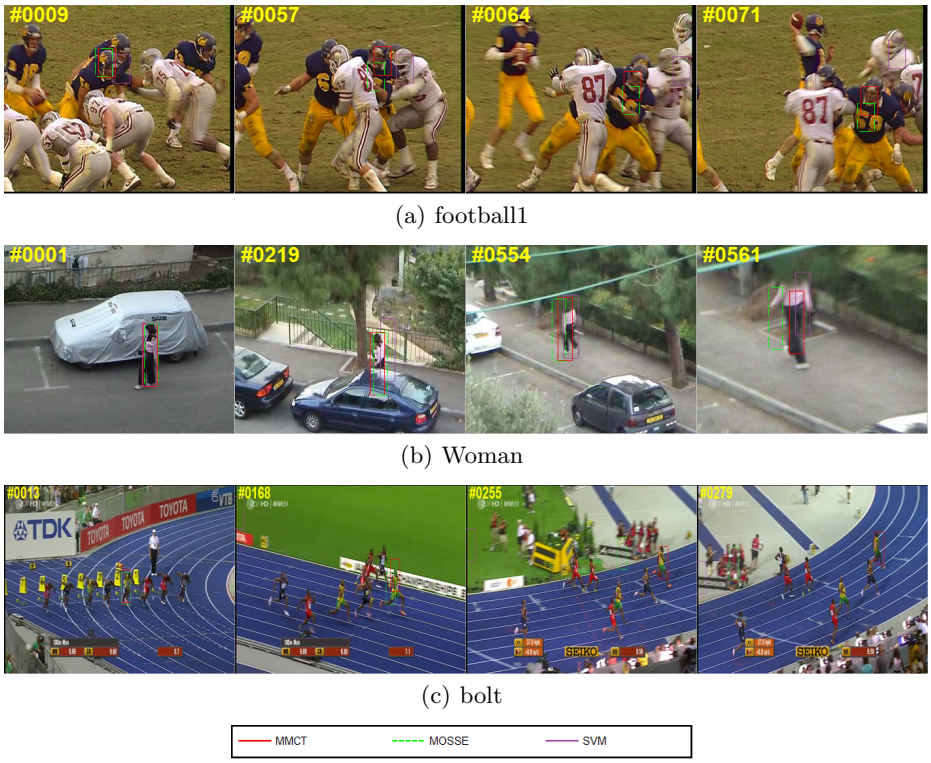


Fig. 3. The tracking results of different trackers

References

1. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. Volume 2., IEEE (2000) 142–149
2. Blake, A., Isard, M., Reynard, D.: Learning to track the visual motion of contours. Artificial Intelligence **78** (1995) 179–212
3. Avidan, S.: Support vector tracking. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Volume 1. (2001) I–184–I–191 vol.1
4. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 2411–2418
5. Avidan, S.: Ensemble tracking. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29** (2007) 261–271
6. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (2009) 983–990
7. Kalal, Z., Matas, J., Mikolajczyk, K.: Pn learning: Bootstrapping binary classifiers by structural constraints. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 49–56
8. Hare, S., Saffari, A., Torr, P.H.: Struck: Structured output tracking with kernels. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 263–270
9. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. (2010) 2544–2550
10. Rodriguez, A., Boddeti, V.N., Kumar, B.V., Mahalanobis, A.: Maximum margin correlation filter: A new approach for localization and classification. Image Processing, IEEE Transactions on **22** (2013) 631–643
11. Bolme, D., Draper, B., Beveridge, J.: Average of synthetic exact filters. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (2009) 2105–2112
12. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. (2012) 702–715
13. Réfrégier, P.: Filter design for optical pattern recognition: multicriteria optimization approach. Optics Letters **15** (1990) 854–856
14. Platt, J., et al.: Sequential minimal optimization: A fast algorithm for training support vector machines. (1998)
15. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust l1 tracker using accelerated proximal gradient approach. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1830–1837
16. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via multi-task sparse learning. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2042–2049