Ann Math Artif Intell (2017) 80:113–129
DOI 10.1007/s10472-017-9543-0

Robust visual tracking using information theoretical learning

Weifu Ding[1,2] · **Jiangshe Zhang**[2]

Published online: 23 March 2017

Abstract This paper presents a novel online object tracking algorithm with sparse representation for learning effective appearance models under a particle filtering framework. Compared with the state-of-the-art ℓ_1 sparse tracker, which simply assumes that the image pixels are corrupted by independent Gaussian noise, our proposed method is based on information theoretical Learning and is much less sensitive to corruptions; it achieves this by assigning small weights to occluded pixels and outliers. The most appealing aspect of this approach is that it can yield robust estimations without using the trivial templates adopted by the previous sparse tracker. By using a weighted linear least squares with non-negativity constraints at each iteration, a sparse representation of the target candidate is learned; to further improve the tracking performance, target templates are dynamically updated to capture appearance changes. In our template update mechanism, the similarity between the templates and the target candidates is measured by the earth movers' distance(EMD). Using the largest open benchmark for visual tracking, we empirically compare two ensemble methods constructed from six state-of-the-art trackers, against the individual trackers. The proposed tracking algorithm runs in real-time, and using challenging sequences performs favorably in terms of efficiency, accuracy and robustness against state-of-the-art algorithms.

Keywords Robust visual object tracking · Information theoretical learning · Adaptive appearance model · Particle filtering · Occlusion and outlier

Mathematics Subject Classification (2010) 68T45

✉ Weifu Ding
dingweifu@163.com

[1] School of Mathematics and Information, BeiFang University of Nationalities, Yinchuan, 750021, China

[2] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China

1 Introduction

Real-time object tracking is an important task in many visual applications including surveillance, augmented reality, medical imaging and driver assistance. Despite significant progress in recent decades, the problem continues to present challenges; it must deal with appearance variations caused by numerous factors including scale, illumination, occlusions, cluttered backgrounds, pose variations, and camera motion.

A tracker generally comprises of three blocks: An image representation that reflects the characteristics of an object's appearance; an effective 2D appearance model that incorporates new information and evaluates the likelihood of a tracking candidate belonging to an object class during tracking; and a search strategy for finding the most likely states in the current frame. Target tracking in the presence of noise and outliers is critically challenging mainly due to the unpredictable nature of errors caused by occlusions, non-Gaussian noise, and large outliers. The corruption may affect any part of the target, and therefore cannot be ignored or treated as minor noise.

Recently, sparse representation has attracted considerable attention in the field of computer vision. Wright et al. [20] proposed a sparse representation classifier for robust face recognition, where an ℓ_1-regularized optimization procedure is adopted to obtain a sparse linear representation solution. The solution has been shown to give state-of-the-art robustness against various disturbances, and particularly occlusions. Some recent research has also focused on using sparse representation for visual tracking [16, 19, 25, 26]. These trackers yield the sparse representation of the target candidate using a dictionary that can be updated gradually; the trackers have demonstrated promising results in various tracking environments, but at the expense of high computational cost largely resulting from ℓ_1 minimization.

Although sparse representation is able to select the most representative templates for each target candidate, it is still not robust enough for contiguous occlusion during visual tracking. Information theoretic learning [17] is a local similarity measure that makes two arbitrary random variables as correlated as possible under the maximum correntropy criterion. Those pixels corresponding to occlusions and outliers in a target candidate will make small contributions to the correntropy between the templates and the target candidate; more emphasis will be given to those pixels corresponding to pixels of the same class as the target candidate. The noise can therefore be handled uniformly within the correntropy framework. By developing a half-quadratic optimization technique and approximately maximizing the objective function in an alternating way, the complex optimization problem is reduced to the learning of a sparse representation, through a weighted linear least squares problem with a non-negativity constraint at each iteration. This can improve recognition accuracy, while the computational cost is much lower than the sparse representation classifier algorithms.

Taking information theoretic learning into account [17], we present an effective appearance model for visual tracking, which has been developed based on the maximum correntropy criterion, along with the ℓ_1 norm penalty. The model is much less sensitive to outliers and can handle occlusion and outliers in a tracked target.

Building on this, the tracking model uses the particle filtering framework. To further improve robustness of our tracker, we dynamically update the target templates and keep the representative templates during the tracking procedure; the template weights are adjusted using the coefficients in the sparse representation. To the best of our knowledge, our proposed method is the first one to combine the maximum correntropy criterion with sparse representation for visual tracking. The learned templates allow for different appearances of the tracked object.

The key contributions of our work are summarized as follows: 1) Our information theoretic learning based visual tracker (ITLT) is effective and sparse for the situations where the tracking target is corrupted by outliers and non-Gaussian noise. In contrast to Euclidean distance, our algorithm adapts a Gaussian kernel function in a principled way. It can automatically detect occlusions and cluttered background, shows robust performance.

2) The earth mover's distance (EMD) is adopted to measure the similarity between a template and tracking result, and compares two images using their color histograms. Templates of the image sequences 'DavidIndoor' are shown in Fig. 1.

The remainder of this paper is organized as follows: In the next section, we review the current state of the art tracking algorithms related to ITLT; in Section 3, we propose ITLT followed by an efficient dictionary update method; the qualitative and quantitative results of numerous experiments and performance evaluations are presented in Section 4. finally, we conclude this paper with remarks on potential future work in Section 5.

2 Related work and motivation

There is a large body of existing research in the field of visual tracking, and an exhaustive discussion of this topic go beyond the scope of this paper. We would refer interested readers to the referenced survey paper [22], and benchmark [21] for a more thorough view. In this section, we briefly review some representative work relating to online object tracking, with emphasis on algorithms that operate directly on grayscale images.

Generally, visual tracking algorithms can be categorized as either generative [1, 6, 8, 16, 18] or discriminative [3, 5, 7, 9, 10], based on their appearance models. Generative methods are centered around a search for the regions that are most similar to the tracked targets, and mainly concentrate on accurately fitting data from the object class. Comaniciu et al. [8] used a histogram computed from a circular region, which was weighted by a spatially smooth isotropic kernel to represent the static template. The tracker embeds the spatially weighted color histogram into a mean shift-based tracking framework and maximizes the appearance similarity iteratively by comparing the histograms of the object and the target candidates. Adam et al [1] constructed a patch-division visual representation with a histogram-based feature description for object tracking; by considering the geometric relationship between patches, it is capable of capturing the spatial layout information.

Subspace representation aims at adapting appearance variations within a low-dimensional subspace, based on the core desire for dimensionality reduction and feature extraction. Black et. al [6] trained an off-line subspace model to represent the object of interest for tracking. In [18], an incremental subspace model (IVT) is used to capture variations in object appearance; the likelihood of a candidate sample belonging to the object class, is often determined by the residual between the candidates and its reconstructed. While the IVT method is effective in handling appearance change caused by illumination variation and pose angle variation, it is not robust with partial occlusion and background clutter. Noisy or misaligned samples are likely to degrade the subspace basis, thereby causing these algorithms to gradually drift away from the target objects.

Fig. 1 Learned templates for faceocc1 sequence. The learned templates cover different appearances of the tracked object

Motivated by the work in [20], sparse representation methods have been used to represent the target, by using a set of target and trivial templates to handle partial occlusion, illumination change and pose variation in visual tracking; the target templates are used to describe the object class to be tracked and trivial templates are used to deal with occlusions and outliers. The likelihood of target candidates is determined by the target templates reconstruction error. Even with further improvements, the ℓ_1 tracker is computationally expensive. The ℓ_1 sparse tracker assumes that the image is corrupted by Gaussian noise, and can not be robust enough for contiguous occlusion, thereby limiting its application in real-time scenarios.

Discriminative methods treat object tracking as a binary classification problem within a local image region, and aim to separate the target object from the background. The process of discriminative methods comprises two key stages: First, in a fixed frame the positive and negative samples are selected to update the online classifiers; Next, in successive frames the target candidates are sampled by the motion model, and the candidate that has highest score in the trained classifier is considered the tracking result. Many sophisticated machine learning technologies were introduced in this framework, and selection of image features plays an important role in the performance of the classification. Collins et al. [7] use the variance ratio of two classes to select discriminative color features for object tracking. Avidan [3] extends a support vector machine classifier within the optical flow framework for object tracking; the tracker aim to learn margin-based discriminative support vector machine (SVM) classifiers for maximizing interclass separability. In [4] and [9], a strong classifier is constructed by selecting several of the most discriminative base classifiers from the Haar-like feature pool. The drawback of this single-instance visual representation is to rely heavily on exact object localization, without which tracking performance may be greatly degraded due to the suboptimal training sample selection. Babenko et al. [5] apply online multiple instance boosting to gain a strong ensemble classifier for object tracking. The tracker representing an object by image patches bag and passing the ambiguity of the samples on to the learning algorithm, can achieve robust tracking results. Unlike existing methods based on classification, Hare et al. [11] proposed an online kernelized structured output SVM for robust tracking, which brings benefits in terms of generalization and robustness to noise; it also shows superior performance compared to state-of-the-art trackers. Furthermore, [12] proposed a fast and robust tracking algorithm that used a circulant kernel matrix structure in the SVM classifier, which can be efficiently computed by the fast Fourier transform. Zhang K [24] adopted a sparse measurement matrix to compress samples and extract the features for the appearance model; the tracking task is then formulated as a binary classification, using a naive Bayes classifier with online updates in the compressed domain.

3 Preliminaries

This section presents some preliminary information regarding the information theoretic learning and particle filtering used in ITLT.

3.1 Particle filtering

The Bayesian approach offers a systematic way of combining prior knowledge of target positions, modeling assumptions, and observation information, to visual tracking [2]. The novel algorithm for a dynamic system can process received data sequentially rather than as

a batch, so that it is unnecessary to store the complete data set, or to reprocess existing data if a new measurement becomes available.

Particle filtering is a Bayesian sequential importance sampling method for estimating the posterior distribution of state variables that characterize a dynamic system. MCMC and online Bayesian methods are adopted to handle the high dimensional complex integral and online processing problems, respectively. Particles are weighted based on a likelihood score, and then propagates these weighted particles according to a motion model. Over the last few years, particle filters have proven to be powerful tools for object tracking, and consist of two steps: Prediction and update. Let $\mathbf{x}_t$ denote the state variables describing the affine parameters of the target at time t; according to the Chapman-Kolmogorov equation, the predicting distribution of $\mathbf{x}_t$ given all available observations up to time $t-1$, is recursively computed as

$$p(\mathbf{x}_t|y_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|y_{1:t-1})d\mathbf{x}_{t-1} \tag{1}$$

The state vector is updated per Bayesian theorem after the observation y_t is available

$$p(\mathbf{x}_t|y_{1:t}) = \frac{p(y_t|\mathbf{x}_t)p(\mathbf{x}_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \tag{2}$$

where $p(y_t|\mathbf{x}_t)$ denotes the observation likelihood. In particle filtering, the underlying posterior $p(\mathbf{x}_t|y_{1:t})$ distribution is approximated by a finite set of N samples$\{\mathbf{x}_t^i\}_{i=1,\ldots,N}$ with important weights w_t^i. The samples $\mathbf{x}_t^i$ are drawn from an importance density $q(\mathbf{x}_t|\mathbf{x}_{1:t-1}, y_{1:t})$, and the weight update equation can then be shown to be

$$w_t^i = w_{t-1}^i \frac{p\left(y_t|\mathbf{x}_t^i\right) p\left(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i\right)}{q(\mathbf{x}_t|\mathbf{x}_{1:t-1}, y_{1:t})} \tag{3}$$

To avoid the degeneracy phenomenon, the particles are resampled to generate a set of equally weighted particles per their important weights. For the calculating convenience, we choose the importance density to be the motion model

$$q(\mathbf{x}_t|\mathbf{x}_{1:t-1}, y_{1:t}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{4}$$

The weights then become the observation likelihood $p(y_t|\mathbf{x}_t)$

3.2 Information theoretic learning

The mean square error (MSE) is probably the most widely methodology for quantifying how similar two random variables are. Novel solutions from MSE rely heavily on the Gaussianity and linearity assumptions. Information theoretic learning (ITL) extracts more information from the data for adaptation under the condition of preserving the nonparametric nature of correlation learning and MSE adaptation; this yields solutions that are more accurate than MSE in non-Gaussian and non-linear signal processing.

Inspired by ITL, Liu et al. [14, 15, 23] recently extended the correlation function for random processes with correntropy to simply estimate directly from statistical samples. The kernel trick is adopted to nonlinearly map the input space to a higher dimensional feature space, and has been shown to obtain robust analysis and efficiently handle non-Gaussian noise and large outliers.

For adaptive systems, the correntropy for any two vectors $A = (a_1, \ldots, a_m)$, $B = (b_1, \ldots, b_m)$ is as follows:

$$\max_{\theta} \frac{1}{m} \sum_{j=1}^{m} g(e_j) \tag{5}$$

where the error defined as $e_j = a_j - b_j$, $j = 1, \ldots, m$, $g(x) = exp(-\frac{x^2}{2\sigma^2})$ is a nonlinear Gaussian kernel function, and θ is the parameter in the criterion to be specified later.

3.3 The ℓ_1 tracker

The location of a target object in a t-th frame can be represented by the six affine transformation parameters $\mathbf{x}_t = (\alpha_t^{(1)}, \alpha_t^{(2)}, \alpha_t^{(3)}, \alpha_t^{(4)}, x_t, y_t)$, where $(\alpha_t^{(1)}, \alpha_t^{(2)}, \alpha_t^{(3)}, \alpha_t^{(4)})$ are the deformation parameters, and (x_t, y_t) denote the translation of the object along the x, y coordinates at time t.

To develop a tracker for generic applications, the state transition equation of the object is modeled by Brownian motion. Each parameter in $\mathbf{x}_t$ is modeled independently by a normal distribution around its counterpart in $\mathbf{x}_{t-1}$, and thus the motion between consecutive frames is itself an affine transformation. Explicitly,

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}, \sigma) \tag{6}$$

where σ is a diagonal covariance matrix whose elements are the corresponding variances of affine parameters. The variances of affine parameters are different and do not change over time. There is a tradeoff between the number of particles needing to be drawn and how well particle filters approximate the posterior distribution; with larger values in the diagonal covariance matrix σ and more particles, it is possible to track the object with higher precision at the price of increased computation.

The particles $\mathbf{x}_t^i, i = 1, \ldots, N$ are found using the particle filter motion model, with each represented by an affine parameter. By applying an affine transformation using $\mathbf{x}_t^i$ as parameters, the i-th target candidate $I_t^i = (y_1, \ldots, y_m)^T$ is cropped from the current frame, and normalized to have the same size as the templates.

In visual tracking, the subspace can be treated as spanned by a set of templates obtained from the previous frame, which consists of n target templates; we denote these as

$$T = [\mathbf{t}_1, \ldots, \mathbf{t}_n] \in R^{d \times n} \tag{7}$$

$\mathbf{t}_i \in R^d$ are column vectors formed by stacking template image columns, t_{ij} is the jth entry of t_i.

For sparse coding of the target candidate, ℓ_1 tracker solves the following optimization problem:

$$min \| \begin{bmatrix} T & E \end{bmatrix} \mathbf{c} - I_t^i \|_2^2 + \lambda \|\mathbf{c}\|_1, s.t.c \succeq 0 \tag{8}$$

Where $\mathbf{c}^T = [\mathbf{a} \quad V]$, E denotes the identity matrix of size $m \times m$, is used to describe the trivial templates, $\mathbf{a}$ indicates the corresponding coefficients, V is the coefficients of trivial templates, and $\| \cdot \|_1$ and $\| \cdot \|_2$ denote the ℓ_1 and ℓ_2 norms, respectively. To solve the above optimization problem, the particle with smallest reconstruction error is chosen as the tracking result.

4 Information theoretic learning tracker

4.1 Sparse representation of a tracked target

It is assumed that the global appearance of an object under a different illumination and viewpoint, lies approximately in a low-dimensional subspace span. In tracking, such a subspace can be treated as spanned by a set of templates T.

Thus, we can represent a target candidate I_t^i by template set T

$$\mathbf{I_t^i} \approx \mathbf{T}\beta = \beta_1\mathbf{t_1} + \beta_2\mathbf{t_2} + \ldots + \beta_n\mathbf{t_n} = \left(\sum_{i=1}^{n} t_{i1}\beta_i, \ldots, \sum_{i=1}^{n} t_{im}\beta_i\right)^T \tag{9}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T$ is called a target coefficient vector.

We wish to find a sparse coding vector $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)^T$ such that $\mathbf{Ta}$ becomes as correlated to $\mathbf{I}_t^i = (y_1, \ldots, y_m)^T \in R$ as possible under the maximum correntropy criterion.

$$\mathbf{a} = \max_{\beta} \sum_{j=1}^{m} g\left(y_j - \sum_{i=1}^{n} t_{ij}\beta_i\right) - \lambda \sum_{i=1}^{n} \beta_i \tag{10}$$

This not only greatly reduces the complexity of the model, but also achieves significantly better performance. To address the computational problem, we first impose nonnegativity constraints on the variables in the correntropy, and then utilize the half-quadratic and EM method to solve the optimization problem (10).

The positive vector $\mathbf{a}$ is actually playing as a clustering indicator, because each entry $\mathbf{a}_i$ reflects the importance of sample $\mathbf{t}_i$ in reconstructing the target candidate I_t^i. Hence, it should be expected that more weight would be assigned to the samples of the same class label I_t^i, while weights of the others should be small or zero in an optimal case. Thus, $\mathbf{a}$ can also be sparse even λ is set to zero. When $\lambda > 0$, ITLT can yield a sparser solution and further improve the recognition accuracy.

After finding the sparse solution to (10), we find the likelihood of each target candidate. The likelihood of candidate targets $p(I_t^i|\mathbf{x}_t^i)$ reflects the similarity between a target candidate and the target templates, and is governed by the correlation between the target candidate and its reconstructed image, based on the information theoretic learning model.

That is the maximal nonlinear difference between the target candidate and its reconstruction.

$$l(I_t^i) = p(I_t^i|\mathbf{x}_t^i) = \sum_{j=1}^{m} g\left(y_j - \sum_{i=1}^{n} \mathbf{t}_{ij}\mathbf{a}_i\right) \tag{11}$$

The ℓ_1 tracker is different, in that assumes that the image pixels are corrupted by independent Gaussian noise. Our robust appearance model considers the effects of occlusion and motion blur, which treat individual pixels of the representation differently, and give more emphasis to those pixels corresponding to pixels of the same class as the target candidate. To capture the appearance variation, our generative appearance model is updated with the tracking result per our mechanism, which updates at each frame.

It should be noted that our object appearance model is different from that used to target templates and trivial templates [16]; in our system, it is unnecessary to include the trivial templates in the templates. The nonlinear kernel function is adopted to assign little weight to the corrupted pixels, and it is this mechanism that can adaptively handle the challenges of a complex environment. Intuitively, only the atoms in target templates which are the same class as the good target candidate will be activated. Similarly, the coefficients corresponding to the atoms that are different with the good target candidate tend to be zero.

The main steps of our algorithm are summarized in **Algorithm** 1.

4.2 Template Update

In practice, object appearance remains the same only for a period of time, after which the template is no longer able to capture the variation in the object appearance.

Algorithm 1 Information learning theory visual tracking

Input: The t-th image frame, and a matrix of dynamic templates parameter T, n_T is the number of templates.

Initialize: Uniformly initialize the weight w to $1/n_T$, and the variances of affine parameters ε.

1: Draw particles according to the dynamical model from the particle filter.

$$\mathbf{x}_t^i \leftarrow \mathbf{x}_{t-1}^i + \varepsilon \tag{12}$$

2: Warp the image with the particle $\mathbf{x}_t^i$, and obtain the candidate image patch $\mathbf{I}_t^i$

3: Solve the information theoretic learning for each I_t^i

4: $p\left(\mathbf{I}_t^i|\mathbf{x}_t^i\right) = \sum_{j=1}^{m} g\left(y_j - \sum_{i=1}^{n} \mathbf{t}_{ij}\mathbf{a}_i\right)$

5: Update the sample weight $w_t^i \leftarrow p\left(\mathbf{I}_t^i|\mathbf{x}_t^i\right)$

6: Find the current target $\mathbf{I}_t^{result}$ which has the largest likelihood

7: Update the templates

Output: Tracking result, and the corresponding parameter to $\mathbf{I}_t^{result}$

4.2.1 Mechanism to detect corruption

We devise a mechanism to detect the degree of pixel corruption, and dynamically update templates $\mathbf{T}$ to address the issue. First, we extracted the weight image of the tracking result. The larger the value of the entry, the more it contributes to the correntropy-base objective function. A fixed constant u is set to judge whether the pixels are occluded; we then count the number of the occluded pixels $N_{occlusion}$ in the weight images which are less than u, and compute the ratio η of the number of occlusion map pixels and the number of target pixels

$$\eta = \frac{N_{occlusion}}{N_{pixel}} \tag{13}$$

Two thresholds w_1 and w_2 are then used to describe the degree of corruption of the image patch. If $\eta < w_1$, the tracking result is directly used to update the template set. If $w_1 < \eta < w_2$, it indicates that the target is partially occluded. We then replace the occluded pixels with corresponding parts of the average observation μ, and use this recovered sample for updates. Otherwise if $\eta > w_2$, it means that a significant part of the target object is occluded, and the templates will not be updated.

4.3 Reducing computation time

The computation load of the proposed algorithm is calculated with the coefficients, using the half-quadratic optimization technique. There are n linear variables corresponding to the number of templates, and m auxiliary variables to the dimension of the template in half-quadratic optimization. In ITLT, it is unnessary to include the trivial template in the dictionary, so the sparse representation computation is extremely large when compared to information theoretic learning; in contrast, all of the m auxiliary variables are updated at each iteration. Thus, ITLT can efficiently estimate the m auxiliary variables in half-quadratic optimization. The ℓ_1 minimization treats $m + n$ equally. When the dimension m is large, the computation cost of sparse representation classifiers will increase rapidly and become

Algorithm 2 Template update

Input: I_t^{result}, β is the newly chosen tracking result and its sparse code,threshold parameter τ, η.
1. **if** $\eta > w_2$
break
Otherwise
2. $\mathbf{w} = \{w_i \leftarrow \|t_i\|_2, i = 1, 2, \ldots, n_T\}$ is the weight vector of current templates.
3. update weights according to the coefficients of target templates $\mathbf{a}_i$, $w_i \leftarrow w_i * exp(\mathbf{a}_i)$
4. $i_0 \leftarrow \mathbf{arg} \min_{1 \le i \le n} \beta_i$
5. if $\mathbf{EMD}\left(I_t^{result}, t_m\right) < \tau, m = \mathbf{arg} \max_{1 \le i \le n} \beta_i$ **&&** $\eta < w_1$
$t_{i_0} \leftarrow \lambda y + (1 - \lambda) t_{i_0}$
else
$t_{i_0} \leftarrow \lambda y + (1 - \lambda) t_{i_0}$
end
6. $w_{i_0} \leftarrow median(w)$
7. Normalize w such that $sum(w) = 1$
end
Output: new template T

extremely expensive. Hence, ITLT is much more efficient than the ℓ_1 tracker and achieves more favorable results in terms of center location error and overlap rate.

5 Experiment

We evaluate ITLT to validate its effectiveness against six state-of-the-art algorithms, namely IVT [18], ℓ_1 [16], FragTrack [1], MILTrack [5], VTD [13], and PN. For the comparison, either the binaries or source codes provided by the authors, with the same initialization and parameter settings were used to generate the comparative results. These sequences involve most challenging situations in visual object tracking, such as heavy occlusions, large pose variations, and drastic illumination changes, as well as low foreground or background contrast. For the trackers involving randomness, we repeat the experiments 10 times on each sequence and report the averaged results.

5.1 Experimental setup

ITLT is implemented in MATLAB, which runs at 2 frames per second on a Pentium 2.0 GHz Dual Core PC, with 3 GB of memory. Table 1 lists the evaluated image sequences; only gray scale information is used for the experiments. The parameters are fixed for all presented sequences. All parameters are set by hand tuning, using some prior knowledge.

The number of target templates used is 40, which is a trade-off between computational efficiency and effectiveness of modeling fast target appearance changes. For particle filtering, 600 particles were used, and the variance σ of observation probability was set to 0.5. Initially, we select the first target template manually from the first frame. The remaining target templates are created by moving a few pixels in four possible directions, at the corner points of the first template in the first frame. All target templates are normalized and weighted uniformly. The overview of ITLT for robust appearance model are shown in Fig. 2.

Table 1 Evaluated video clip

Video Clip	Frames	Challenging factors
Car4	659	illumination variation, scale change
David Indoor	462	illumination variation , scale change,out-plane rotation
Occlusion 1	898	partial occlusion
Occlusion 2	819	occlusion in-plane rotation,out-plane rotation
Stone	593	partial occlusion,background clutter
Singer	321	illumination variation,scale change

5.2 Qualitative comparison

Pose and illumination change In the Car4 sequences, the tracked vehicles are moving on an open road, and the target undergoes drastic illumination changes as it passes beneath a bridge and under trees. Since the target is a rigid object, its shape and scale does not change greatly. Some samples of the final tracking results are shown in Fig. 3a; the frame indices are 1, 150, 200, 249, 548 and 659. Generative methods like IVT, and ℓ_1 tracker perform well for this sequence. From Fig. 3a, we can see that our tracker is capable of constantly tracking the car, even if the illumination changes drastically.

For the DavidIndoor sequence shown in Fig. 3b, the appearance changes gradually due to illumination, pose and scale variation when the person shown walks out of a dark meeting room; out-of-plane rotation also occurs in some frames. Our tracker, IVT, VTD and TLD algorithms can accurately locate the true target without great offset on this sequence. The IVT method uses a PCA-based appearance model that has been shown to accurately account for appearance change caused by illumination variation. The VTD method performs well, due to the use of multiple observation models constructed from different features. The TLD approach also works well, because it maintains a detector that uses Haar-like features during tracking. Our tracker uses the templates updated by subspace learning, and can efficiently capture the variation in the tracked object.

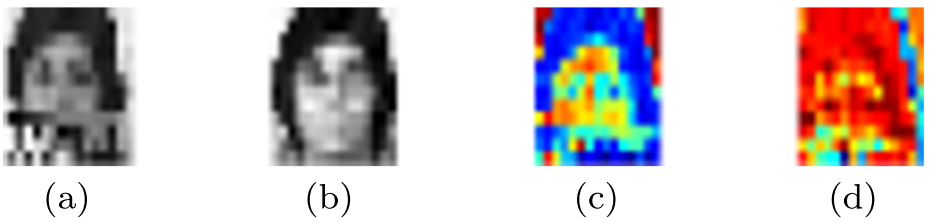

Fig. 2 Overview of our tracker for robust appearance model. **a** A target candidate cropped from current with book occlusion. **b** The reconstructed image by a learned sparse linear combination of all of the training images. **c** The reconstructed error. **d** The weight image learned by our appearance model. The entry with blue color has a small value, while the entry with red color has a large value. The larger the value of the entry, the more it contributes to the correntropy-based objective function. Due to the occlusion caused by the book,the pixels under the mouth are assigned small weights, which means that they are estimated as noise. **e** The sparse coefficients computed by our approach. The red coefficients correspond to the template with the similar appearance as target candidate

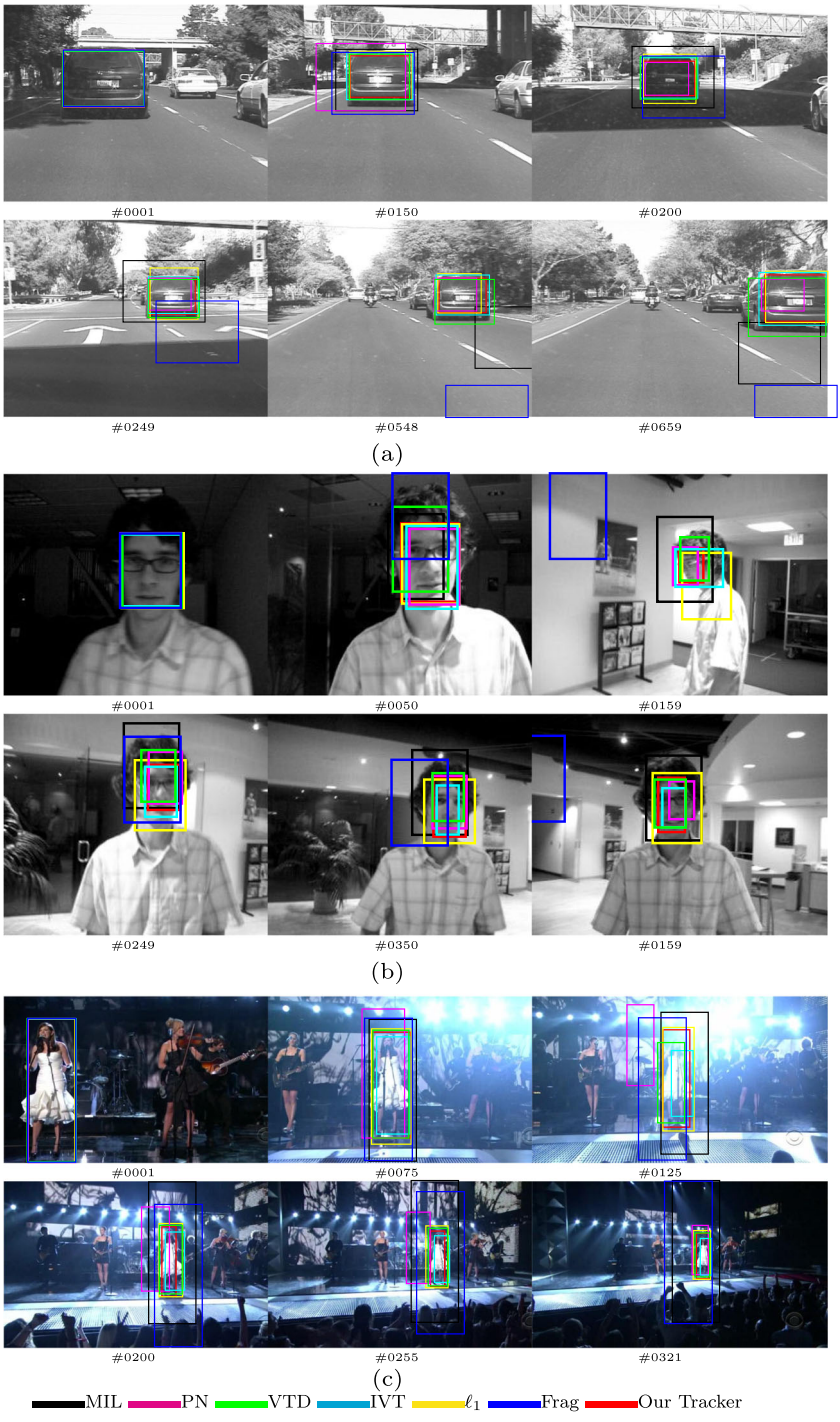

Fig. 3 Qualitative evaluation:object appearance change drastically due to large variation of lighting, pose, scale. **a** Car4. **b** Davidindoor. **c** Singer

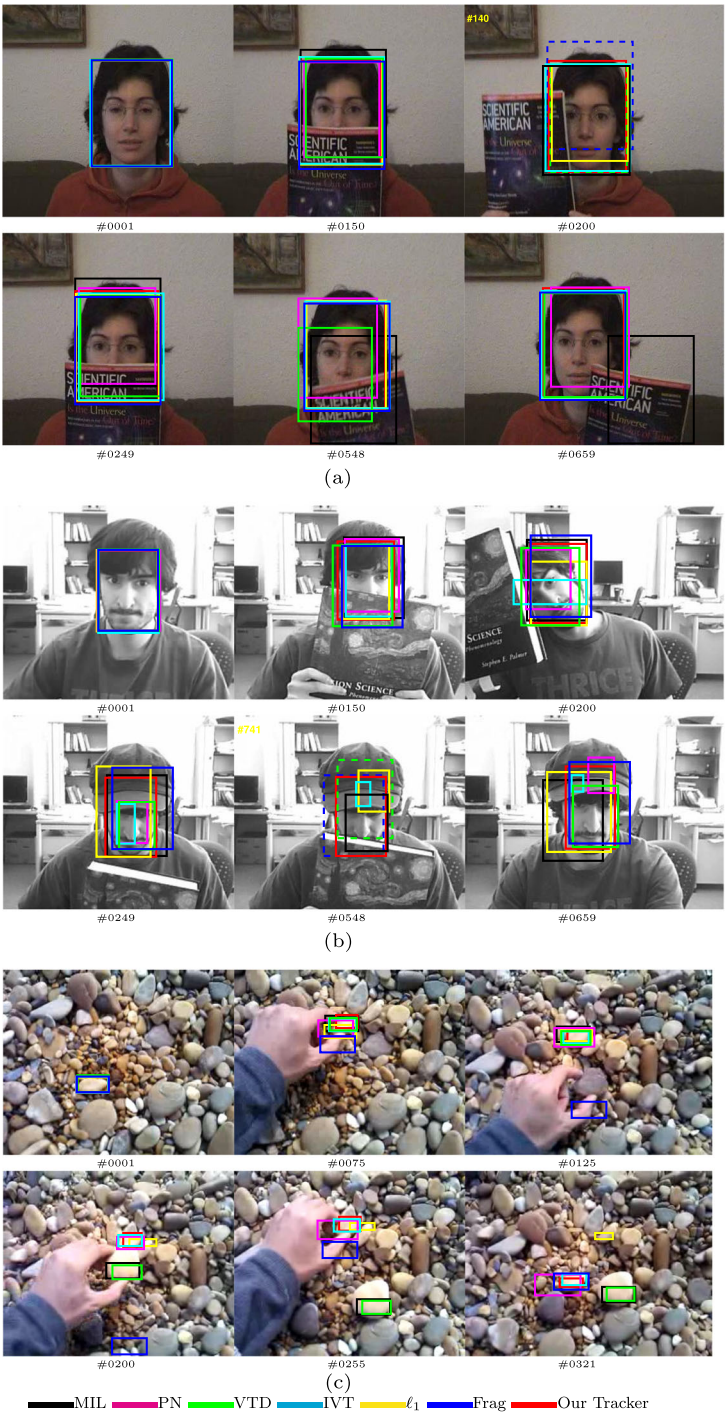

Fig. 4 Qualitative evaluation: object appearance change drastically due to large variation of occlusion, clutter background, lighting. **a** faceocc1. **b** faceocc2. **c** stone

In the Singer1 sequence shown in Fig. 3c, the singer progressively undergoes large pose and illumination changes. The MIL methods perform well on this sequence with lower tracking errors than other methods. The IVT, and l1 methods do not perform well on this sequence because they use holistic features that are less effective for large scale pose variations. The features used in the proposed algorithms are similar to generalized Haar-like features, which have been shown to be robust to pose and orientation change.

Occlusion and pose variation The target objects are partially occluded in the faceocc1 and faceocc2 sequences, and six representative frames are shown in Fig. 4a and b. Most tracking methods do not perform well when the objects are heavily occluded. In the faceocc1 sequence, our tracker, FragTrack and the ℓ_1 tracker perform better, as shown in Fig. 4a; this is because these methods take partial occlusion into account. The IVT tracker drifts to the un-occluded face region. For the faceocc2 sequence, the FragTrack method performs poorly since it does not handle appearance change caused by pose and occlusion. Although the MIL is able to track the target object, it is unable to estimate the in-plane rotation due to its design. By assigning smaller weights to the pixels around the occlusions, our tracker performs best, particularly when partial occlusion or in-plane rotation occurs.

Background clutter In the Stone sequence, the target objects undergo fast movement in cluttered backgrounds; additionally, the target and surrounding background are similar to the target object. The ℓ_1 and Fragtracker perform well because the surrounding background is similar to the target object. The IVT fails after an abrupt motion occurs, and the PN tracker drifts gradually. Both the MIL and our proposed algorithm are able to track the right objects accurately; in the case of our algorithm, the result of its robust mechanism for addressing occlusions and outliers. The proposed algorithm also adapts better to change of rotation.

5.3 Quantitative comparison

Two common performance metrics for quantitative comparison are used to evaluate the proposed algorithm with 6 state-of-the-art trackers. Gray scale videos are used. The first metric is the success rate which is defined in the PASCAL VOC challenge as score$=\frac{area(ROI_T \bigcap ROI_G)}{area(ROI_T \bigcup ROI_G)}$, where ROI_T is the tracking bounding box and ROI_G is the ground truth bounding box. If the score is larger than 0.5 in one frame, the tracking result is considered a

Table 2 Average center location error (pixels)

Video Clip	MIL	Frag	PN	VTD	IVT	L1	Ours
car4	60	180	13	13	3	5	6
davidindoor	14	75	9	11	6	10	5
faceocc1	31	8	13	8	7	7	7
faceocc2	12	14	15	11	10	10	9
stone	33	68	9	33	4	20	7
singer1	15	21	33	4	8	4	6
Average	27.5	61	15.3	13.3	6.3	9.3	6.7

Table 3 Overlap rate tracking methods

Video Clip	MIL	Frag	PN	VTD	IVT	L1	Ours
car4	0.30	0.20	0.64	0.67	0.82	0.78	0.87
davidindoor	0.41	0.18	0.60	0.56	0.68	0.54	0.79
faceocc1	0.60	0.82	0.68	0.79	0.84	0.86	0.94
faceocc2	0.71	0.65	0.50	0.60	0.53	0.70	0.88
stone	0.34	0.15	0.38	0.38	0.62	0.31	0.78
singer1	0.36	0.35	0.41	0.76	0.61	0.72	0.88
Average	0.45	0.39	0.54	0.63	0.68	0.65	0.86

success. The other metric is the center location error, which is defined as the Euclidean distance between the central locations of the tracked objects, and the manually labeled ground truth.

Tables 2 and 3 show the tracking performance of our method with the 6 other methods. We note that the TLD tracker does not report a tracking result when the drift problem occurs and the target object is redetected. Thus, we only report center location errors for the sequences in which the TLD method does not lose track of target objects. The proposed tracker performs favorably against the state-of-the-art algorithms, achieving best or second-best performance in most sequences using both evaluation criteria. Figure 5 some shows center error tracking results for different trackers. The center location error of our tracker is much smaller than those of the other trackers, and as shown in Table 4, our tracker achieves higher success rates than the others. This demonstrates the advantage of our proposed algorithm. Results showing the overlap rate for different trackers are shown in Fig. 6.

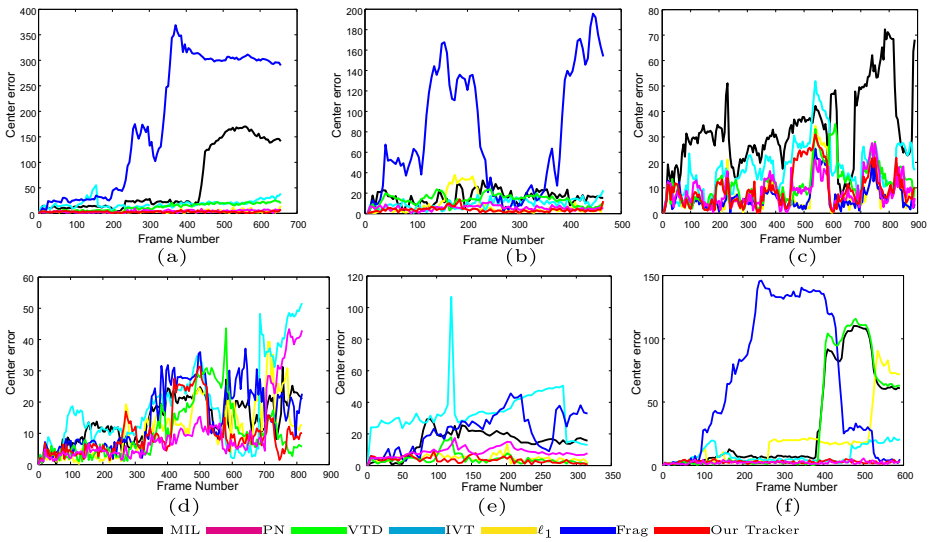

Fig. 5 Tracking results of the center error. The figure shows center error for six video clips we tested on. Our algorithm is compared with six state-of-the-art methods: IVT, ℓ_1 tracker, FragTrack, MILTrack, VTD and PN methods. **a** car4. **b** davidindoor. **c** faceocc1. **d** faceocc2. **e** singer. **f** stone

Table 4 Success rate of tracking

	MIL	Frag	PN	VTD	IVT	L1	Ours
car4	0.24	0.23	0.88	0.93	0.99	0.99	0.96
davidindoor	0.18	0.05	0.71	0.74	0.90	0.59	0.97
faceocc1	0.78	0.99	0.78	0.99	1	1	0.97
faceocc2	0.90	0.87	0.51	0.68	0.52	0.92	0.94
stone	0.32	0.18	0.11	0.43	0.87	0.37	0.92
singer1	0.26	0.27	0.45	0.95	0.62	0.97	0.95
Average	0.45	0.43	0.57	0.79	0.82	0.81	0.95

5.4 Discussion

The experiments demonstrate the robust tracking performance of our algorithm. We note that sparsity, and robustness to occlusions and outliers, is the prime characteristic of information theoretic learning. The robust tracking performance of our algorithm can be attributed to several factors. One of these is that our information theoretic learning model can adaptively assign weight to the pixels; the pixels around the occlusion and outlier are assigned small weights, which means that they are estimated as noise. With fixed poses, the appearance of an object under different illumination conditions can be accurately approximated by a low dimensional subspace. In addition, we have devised a mechanism to detect corruption and update template; this can capture the variation of the target object, and efficiently improve the visual tracking performance.

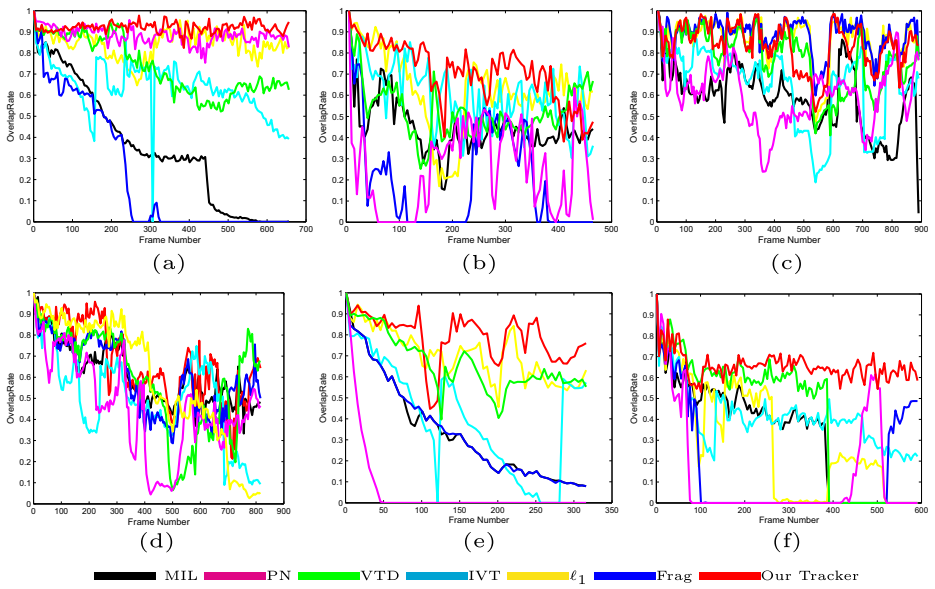

Fig. 6 Tracking results of the overlap rate. The figure shows overlap rates for six video clips we tested on. Our algorithm is compared with six state-of-the-art methods: IVT, ℓ_1 tracker, FragTrack, MILTrack, VTD and PN methods. **a** car4. **b** davidindoor. **c** faceocc1. **d** faceocc2. **e** singer. **f** stone

6 Conclusions and future work

In this paper, we proposed a robust tracking algorithm with a dynamic online updated sparse dictionary, which can adapt to occlusions and outliers. The target appearance is modeled using a sparse representation based on an information theoretic learning model. The natural combination of a dynamic basis and adaptive assignment of weights to different pixels provides a robust sparse appearance model for tracking. To our knowledge, this is the first time that occlusions and outliers have been solved simultaneously, and experimental results demonstrate the effectiveness of the method. In the future, we will extend our representation scheme for other visual problems, including object recognition; we will also develop other maximum correntropy criterion methods using the proposed model.

Acknowledgments This work was supported by the National Basic Research Program of China (973 Program) under Grant no. 2013CB329404, the Major Research Project of the National Natural Science Foundation of China under Grant no. 91230101, the National Natural Science Foundation of China under Grant no. 61075006 and 11201367, the Key Project of the National Natural Science Foundation of China under Grant no. 11131006 and the Research Fund for the Doctoral Program of Higher Education of China under Grant no. 20100201120048, natural science Fund of Ningxia, China under Grant no. NZ12209.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 798–805 (2006)
2. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
3. Avidan, S.: Support vector tracking. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1064–1072 (2004)
4. Avidan, S.: Ensemble tracking. Proceedings of the 10th European Conference on Computer Vision, 494C501 (2005)
5. Babenko, B., Yang, M., Belongie, S.: Visual tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1619–1632 (2011)
6. Black, M.: EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. Int. J. Comput. Vision **26**(1), 63C84 (1998)
7. Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1631–1643 (2004)
8. Comaniciu, D., Member, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003)
9. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. Proceedings of the British Machine Vision Conference, 47–56 (2006)
10. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. Proceedings of the 10th European Conference on Computer Vision, 234–247 (2008)
11. Hare, S., Saffari, A., Torr, P.: Struck:structured output tracking with kernels. Proceedings of the International Conference on Computer Vision and Pattern Recognition, 263–270 (2011)
12. Henriques, J., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. Proceedings of European Conference on Computer Vision, 702–715 (2012)
13. Kwon, J., Lee, K.: Visual tracking decomposition. Proceedings of the International Conference on Computer Vision and Pattern Recognition, 1269–1276 (2010)
14. Liu, W., Pokharel, P., Principe, J.: Error Entropy, Correntropy and M-Estimation. Proceedings Workshop of Machine Learning for Signal Processing (2006)
15. Liu, W., Pokharel, P., Principe, J.: Correntropy: Properties and applications in Non-Gaussian signal processing. IEEE Trans. Signal Process. **55**(11), 5286–5298 (2007)
16. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2259–2272 (2011)
17. Ran, H., Zheng, S., Gang, H.: Maximum correntropy criterion for robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1561–1576 (2011)

18. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. **77**(8), 125–141 (2008)
19. Wang, D., Lu, H., Yang, M.: Online object tracking with sparse prototypes. IEEE Trans. Image Process. **22**(1), 314–325 (2013)
20. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Roubust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009)
21. Wu, Y., Lim, J., Yang, M.: Online object tracking: A benchmark. Proceedings of the International Conference on Computer Vision and Pattern Recognition (2011)
22. Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A survey. ACM Comput. Surv. **38**(4), 81–93 (2006)
23. Yuan, X., Hu, B.: Robust Feature Extraction via information theoretic learning. Proceedings of International Conference on Machine learning, 1193–1200 (2009)
24. Zhang, K., Zhang, L., Yang, M.: Real-time compressive tracking. Proceedings of the 10th European Conference on Computer Vision, 864–877 (2012)
25. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Low-rank sparse learning for robust visual tracking. Proceedings of the 10th European Conference on Computer Vision, 470–484 (2012)
26. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via multi-task sparse learning. Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2042–2049 (2012)