http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (9): 1727-1736, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.1727.1736


# Visual Tracking Based on Sparse Representation in a Co-training Framework

Hanling Zhang and Fei Tao
School of Information Science and Engineering, Hunan University, Changsha, 410082, China

**Abstract:** Visual tracking is an important topic in the field of computer vision and artificial intelligence. The main challenging issue in designing a robust tracking algorithm is the appearance variations caused by numerous factors such as occlusion, background clutter, illumination change and motion blur. In recently years, sparse representation has been extensively studied and applied in visual tracking. The representation has been shown to be robustness to a wide range of image corruptions, especially to an occlusion. However, sparse coding based trackers at a computational expense of the L1 minimization. In this study, we present a novel tracking method based on sparse representation in a co-training framework, exploiting the strength of both holistic representation and local histogram. We first introduce $l_1$ regularization into subspace representation with Principal Component Analysis (PCA). Then, we develop a novel histogram-based tracking method in which we take the spatial information of patches into consideration with an occlusion handling mechanism. Furthermore, we combine them in a novel collaborative model and the update scheme can make the tracker deal with appearance effectively and alleviate the impact of the drift problem. Experimental results on benchmark challenging sequences demonstrate that the robustness and effectiveness of the proposed algorithm is competitive to the state-of-the-art tracking methods.

**Key words:** Visual tracking, sparse representation, appearance model, update scheme

## INTRODUCTION

Despite great progresses in the past decades, visual tracking remains a challenging problem as robust tracking algorithms entail the need to account for appearance variation caused by occlusion, illumination change, background clutter, pose variation, motion blur and rotation. Moreover, few attempts have been made to directly solve the occlusion problem, which remains as arguably the most critical factor for causing tracking failures.

To account for the appearance variations of the target, visual tracking problem has been formulated in two different categories: generative and discriminative. Generative methods typically learn an appearance model to describe the target observations and then use it to search for the regions within the highest probability. Black and Jepson (1998) learn a subspace model offline to represent the target objects for tracking. Matthews *et al.* (2004) propose a template update method which can reduce the drifting problem by aligning with the first template to reduce drifts. Later, visual tracking via online subspace learning has attracted more and more attention. Ross *et al.* (2008) present a tracking method with incremental subspace learning, in this representation, the Sequential Karhunen Loeve (SKL) algorithm is extended to effectively learning the variations of both illumination and appearance. Kwon and Lee (2010) propose a tracking algorithm with the visual tracking decomposition scheme. In this scheme, the observation model is decomposed into multiple basic models to cover a wide range of poses and illumination changes. For discriminative methods, tracking is treated as a classification problem which aims at designing a classifier to effectively separate the foreground target from the background. Collins *et al.* (2005) proposed an online feature selection method to select the most discriminative color space for tracking. Grabner *et al.* (2006) presented an online boosting algorithm to select discriminative features for visual tracking and later a classifier within the semi-supervised learning framework (Grabner *et al.*, 2008) is adapted to address the online update problem. Avidan (2007) extend a Support Vector Machine (SVM) classifier within the optical flow framework for object tracking. Kalal *et al.* (2010) propose the P-N learning algorithm. They exploit the underlying structure of positive and negative samples to learning effective classifier for visual tracking. Babenko *et al.* (2011) introduced Multiple Instance Learning (MIL) into online tracking, where positive and negative samples are put into bags or sets to learn a

**Corresponding Author:** Fei Tao, School of Information Science and Engineering, Hunan University, Changsha, 410082, China

discriminative model. In Wang *et al.* (2011), a discriminative appearance model based on superpixels is presented, thereby facilitating a tracker to separate the foreground target from the background. While these discriminative methods perform well, they nevertheless do not take correctly labeled samples into account which can be useful in updating the classifier.

Recently, researchers have introduced sparse representation for visual tracking (Mei and Ling, 2009) and it is solved through a series of L1 minimization problems to solve the model tracking problem. The method demonstrates promising robustness to a wide range of object corruptions, especially partial occlusion. However, the algorithm with L1 minimization formulation at the expense of high computational cost and it also neglects the local visual information, which will result in bad tracking performance in cases of there is similar object or heavy occlusion. Zhang and Liu (2013) proposed an object detection algorithm for visual tracking, they adopt the variant of the Douglas-rachford Splitting Method (VDRSM) to restore background and foreground by taking advantage of the separable structure.

In this study, we propose a robust visual tracking algorithm in a co-training framework. The proposed object appearance model exploits the strength of both holistic representation and local histogram. The proposed tracking method is effective in dealing with appearance changes through incremental subspace learning and the computation complexity is reduced. In addition, the developed update scheme considers whether the target object is occluded or not, thereby enabling the tracker to deal with appearance change effectively. Experimental results on several challenging sequences demonstrate the robustness and effectiveness of the proposed algorithm, especially when the objects exhibit large appearance changes.

## OBJECT TRACKING WITH SPARSE REPRESENTATION

The pioneering work on applying sparse representation to object tracking is done by Mei and Ling (2009), who proposed a L1 tracker by casting the problem as determining the most likely patch with a sparse representation of object templates and modeling partial occlusion by sparse representation of trivial templates:

$$y = Dx + e = [D \;\; I]\begin{bmatrix} x \\ e \end{bmatrix} = Bw \tag{1}$$

where $y \epsilon R^d$ denotes an observed target sample (by stacking columns to from a 1D vector), $D = [d_1, d_2,..., d_n] \epsilon R^{d \times n}$ ($f >> n$) indicates a set of training templates, $x = (x_1, x_2,..., x_n)^T$ is the corresponding target coefficient vector, $e \epsilon R^d$ is a sparse error and nonzero entries correspond to pixels in y that are occluded or corrupted and $I = [i_1, i_1, ..., i_n]$ is an identity matrix, where each trivial template $i_k$ is a vector with only one nonzero entry in the k-th position.

When the vector w is sparse enough the target coefficients x and sparse error e can be jointly solved by the following $l_0$-norm minimization:

$$w = \arg\min_w \|w\|_0 \qquad \text{s.t. } \| y - Bw \|_2 < \varepsilon \tag{2}$$

where $\|.\|_0$ denotes the $l_0$-norm that counts the number of nonzero entries in a vector, $\|.\|_2$ denotes the $l_2$-norm (i.e., Euclidean distance) and $\varepsilon > 0$ is the noise level. The problem (2) is in general ill-posed (NP-Hard) and has index complexity of the algorithm in theory or practice. To make the problem tractable, the $l_1$-norm minimization is widely used to replace the $l_0$-norm minimization:

$$w = \arg\min_w \|w\|_1 \qquad \text{s.t. } \| y - Bw \|_2 < \varepsilon \tag{3}$$

The Lagrangian version of the problem (3) can be written as:

$$w = \arg\min_w \frac{1}{2} \|y - Bw\|_2^2 + \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} \|w\|_2^2 \tag{4}$$

where, $\lambda_1$ and $\lambda_2$ are regularization parameters, which control the relative importance of the sparseness to the reconstruction error. When $\lambda_2 = 0$, it leads to the widely $l_1$-norm minimization problem. When $\lambda_2 > 0$, it makes the problem (4) become strictly convex. After obtaining the sparse coefficients w, the input sample y can be represented in a sparse way.

## PROPOSED TRACKING FRAMEWORK

We first describe how the holistic and local visual information are exploited. Then the collaborative model and the update scheme of our appearance model are presented.

## HOLISTIC REPRESENTATION

Motivated by above-mentioned discussions, in this study, we combine the incremental subspace learning with sparse representation for modeling object appearance. We first model the target appearance with PCA basis vectors and handling partial occlusion with trivial templates by:

$$y = Ux + e = [U \quad I]\begin{bmatrix} x \\ e \end{bmatrix} \tag{5}$$

where, y denotes an observation vector, U represents a matrix of column basis vectors, x indicates the corresponding coefficients and e is the error term which can be viewed as the coefficients of trivial templates.

By assuming that each candidate image is sparsely represented by a set of target and trivial templates, and error e can be modeled by arbitrary but sparse noise, Eq. 5 can be solved via $l_1$-norm minimization:

$$\min_{x,e} \frac{1}{2}\|y - Ux - e\|_2^2 + \lambda\|e\|_1 \tag{6}$$

Here, Eq. 6 is the variant of Eq. 4 when $\lambda_2 = 0$. Our formulation maintains the holistic appearance information and has the following advantages. On the one side, overcoming the drawback of the incremental subspace representation of the IVT method (Ross *et al*., 2008) is sensitive to partial occlusion, our method handles partial occlusion with trivial templates explicitly. On the other side, comparison with the time complexity of the $l_1$ method (Mei and Ling, 2009) is quite significant, our algorithm is able to reduce the computational complexity by exploiting subspace representation.

Let the object function be $W(x,e) = \arg\min_{x,e} \frac{1}{2}\|y - Ux - e\|_2^2 + \lambda\|e\|_1$

the optimization problem can be constructed as:

$$\min_{x,e} \quad W(x,e) \qquad \text{s.t.} \quad U^TU = I \tag{7}$$

where, I denotes an identify matrix. For solve this optimization problem, an iterative algorithm composes of a simple least squares problem and a shrinkage operation is presented, as shown in Algorithm 1. Thus, the confidence value $L_1$ (where i denotes the i-th sample) can be measured by the reconstruction error of each observed image patch:

$$L_i = \exp(-\|y^i - Ux^i\|_2^2) \tag{8}$$

In order to further deal with occlusions, we use a mask to factor out non-occluding and occluding parts:

$$L_i = \exp\left[-(\|\rho^i \odot (y^i - Ux^i)\|_2^2 + \beta\sum(1 - \rho^i))\right] \tag{9}$$

where, $\odot$ denotes the element-wise multiplication, $[\rho^i_1, \rho^i_2, , ..., \rho^i_d]^T$ is a vector that indicates the zero elements of error $e^i$ and $\beta$ is a penalty term (simply set to $\lambda$ in this study). If the j-th element of $e^i$ is zero, then $\rho^i_j = 1$, otherwise $\rho^i_j = 0$ The first part of the exponent accounts for the reconstruction error of un-occluded portion of the target object and the second term aims to penalize labeling any pixel as being occluded. The experimental results on sequences Caviar1 and Occlusion2 with severe occlusions in a later section demonstrate the effectiveness of our formulation.

Algorithm 1: For computing $e^{opt}$ and $x^{opt}$

Input: An observation vector y, orthogonal basis vectors U and a regularization parameter$\lambda$.

1. Initialize $e^0 = 0$ and i = 0.
2. Iterate
3. Obtain $x^{i+1}$ via $x^{i+1} = U^T (y\text{-}e^i)$
4. Obtain $e^{i+1}$ via $e^{i+1} = S_\lambda (y\text{-}Ux^{i+1})$
5. i←i+1
6. Until convergence or termination

Output: $e^{opt}$ and $x^{opt}$

## LOCAL HISTOGRAM

In this part, sparse codes of local patches with spatial layout in a target object are used to model the appearance model for visual tracking and this can help locate the target more accurately. We first extract a set of overlapped local image patches within a target region and turn them into vectors as $y = [y_1, y_2, ...y_N] \in R^{G\times N}$, where N is the number of local image patches and G denotes the size of each patch. The sparse coefficient $\alpha_i$ corresponds to $y_i$ can be computed by:

$$\alpha = \arg\min_{\alpha_i} \|y_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{10}$$

where, the dictionary $D \in R^{G\times M}$ is generated from k-means cluster centers via the patches belonging to the labeled object in the first frame and M denotes the number of cluster centers. Then, the sparse coefficient vector $\alpha_i$ of each patch is concatenated to form a histogram as $h = [\alpha_1, \alpha_2, ... \alpha_N]^T$.

However, the constructed histogram does not consider partial occlusion. Thus, we regard the local patch with large reconstruction error as corruption and the corresponding sparse coefficient is set to 0. Then, a weighted histogram can be constructed as:

$$\varphi = h \odot o \tag{11}$$

where, each element of o is an indicator of corruption of the corresponding patch and is determined by:

$$o_i = \begin{cases} 1 & \varepsilon_i < \varepsilon_0 \\ 0 & \text{ortherwise} \end{cases} \tag{12}$$

where, $\varepsilon_i = \|y_i\text{-}D\alpha_i\|_2^2$ is the reconstruction error of the local patch $y_i$ and $\varepsilon_o$ is a predefined threshold. The similarity of histograms between the candidate and the

model can be computed by using the histogram intersection function due to its effectiveness:

$$H_i = \sum_{j=1}^{M\times N} \min(\phi_i^j, \varphi^j) \quad (13)$$

where, $\phi_i$ is the histogram for the i-th candidate and $\varphi$ is the template.

In our approach, the reconstructed coefficient of each local patch essentially represents the importance of each local image patch. In the proposed representation mechanism, spatial information of local patches and occlusion are taken into account through which the appearance model is more effective and robust.

## COLLABORATIVE MODEL

In this study, we embed the appearance models into the particle filtering framework to form a robust tracking algorithm and our approach uses the collaborative strength of both holistic representation and local histogram. The particle filtering (Li *et al*., 2008) is a sequential Monte Carlo method, which recursively approximates the posterior distribution characterizing a dynamic system. Given the observation set of the target $y_{1:t} = \{y_1, y_2, ..., y_t\}$ up to time t, the observation likelihood of i-th candidate at time t can be measured by:

$$p(y_t^i \mid x_t^i) = \max(w_i^1 \times L_i, w_i^2 \times H_i) \quad (14)$$

$$w_i^1 = \frac{L_i}{L_i + H_i} = \frac{\exp\left[-(\|\rho^i \odot (y^i - Ux^i)\|_2^2 + \beta\sum(1-\rho^i))\right]}{\exp\left[-(\|\rho^i \odot (y^i - Ux^i)\|_2^2 + \beta\sum(1-\rho^i))\right] + \sum_{j=1}^{M\times N}\min(\phi_i^j, \varphi^j)}$$

$$w_i^2 = \frac{H_i}{L_i + H_i} = \frac{\sum_{j=1}^{M\times N}\min(\phi_i^j, \varphi^j)}{\exp\left[-(\|\rho^i \odot (y^i - Ux^i)\|_2^2 + \beta\sum(1-\rho^i))\right] + \sum_{j=1}^{M\times N}\min(\phi_i^j, \varphi^j)} \quad (15)$$

In the particle filtering framework, the tracking process is governed by a dynamic model $p(x_t|x_{t-1})$ and an observation model $p(_{YT}|x_t)$. The dynamic model $p(x_t|x_{t-1})$ represents the motion state of a target in consecutive frames. Then, we apply an affine image warp to model the target motion state, i.e., $p(x_t|x_{t-1}) = N(x_t; x_{t-1}, \Psi)$ where $\Psi$ is a diagonal covariance matrix whose elements are the variances of the affine parameters. Therefore, the posterior probability $p(x_t|y_{1:t})$ can be inferred by the Bayesian inference recursively:

$$p(x_t \mid y_{1:t}) \propto p(y_t \mid x_t)\int p(x_t \mid x_{t-1})p(x_{t-1} \mid y_{1:t-1})dx_{t-1} \quad (16)$$

Finally, the optimal state $\hat{x}_t$ of the tracked target is obtained by the Maximum a Posteriori (MAP) estimation:

$$\hat{x}_t = \arg\max_{x_t} p(x_t \mid y_{1:t}) \quad (17)$$

## UPDATE SCHEME

The appearance of an object may change drastically due to the inevitable challenging factors. Therefore, an effective online update scheme is important and necessary. In this paper, we present an update scheme in which the holistic and local appearance models are updated independently.

For holistic representation, we employ one of the three kinds of operations based on the occlusion ratio $\eta$ (i.e., the ratio of the number of nonzero pixels and the number of occlusion map pixels). Here, two thresholds $t_{r1}$ and $t_{r2}$ represent the degree of occlusion. First, if $\eta<t_{r1}$, we directly update the model with the sample. Second, if $\eta>t_{r2}$, it means that the target is partially occluded. We then replace the occluded pixels by its corresponding parts of the average observation and use this recovered sample for updating. Third, if $\eta>t_{r2}$, it indicates that a significant part of the target object is occluded and we discard this sample without updating. After we cumulate enough samples, we use the incremental PCA scheme (Ross *et al*., 2008) to update our observation model. With this update strategy, the model can adapt to the appearance change of the target and handle partial occlusion.

For local histogram, in order to capture the new appearance and recover the object from occlusions, the histogram is updated by:

$$\Psi_n = \mu\psi_f + (1-\mu)\psi_l \text{ , if } O_n < O_0 \quad (18)$$

where, $\Psi_n$ indicates the new obtained histogram, $\Psi_f$ denotes the histogram at the first frame and $\Psi_l$ denotes last stored according to the weights assigned by the constant $\mu$. The variable $O_n$ denotes the occlusion condition and can be computed by the corresponding occlusion indication vector $o_n$ (by Eq. 12) using:

$$O_n = \sum_{j=1}^{M\times N}(1 - o_n^i) \quad (19)$$

In this way, the proposed update scheme not only effectively alleviates the visual drift problem, but also captures the variations of the target during the tracking process.

## EXPERIMENTAL RESULTS

Here, we evaluate our tracker on nine challenging image sequences (most of them are publicly

available). The challenging factors in these sequences include occlusion, illumination variation, background clutter, motion blur and rotation. For comparison, we evaluate the proposed tracker against six state-of-the-art algorithms, including FragTrack (Frag) (Adam *et al*., 2006), the Incremental Visual Tracking (IVT) (Ross *et al*., 2008), L1 tracking ($l_1$) (Mei and Ling, 2009), Multiple Instance Tracking (MIL) (Babenko *et al*., 2009), PN learning tracking (PN) (Kalal *et al*., 2010) and Visual Tracking Decomposition (VTD) (Kwon and Lee, 2010) methods. Both qualitative and quantitative evaluations on benchmark challenging sequences demonstrate the favorable performance of the proposed tracking method.

In our experiments, the location of the target object is manually labeled in the first frame for each sequence, the regularization constant $\lambda$ is set to 0.01 ($\beta$ is the same) and the number of particles is set to 600. For PCA representation, each image observation is normalized to 32×32 pixels and 16 eigenvectors are used in all experiments. In addition, 1024 trivial templates are used in this paper. The threshold $\varepsilon_0$ in Eq.12 is set to 0.1. The constant $\mu$ is set to 0.95 and the threshold $O_0$ in Eq.18 is 0.8. Two thresholds $t_{r1}$ and $t_{r2}$ are set to 0.1 and 0.6, respectively.

## QUALITATIVE EVALUATION

Figure 1a and b, respectively show the tracking results on sequences Caviar1 and Occlusion2 with severe occlusions. In object tracking, occlusion is one of the most challenging problems and it is the critical factor causing drift. In the Caviar1 sequence, some trackers fail after heavy occlusion (e.g., #154, #190, #256 and #382). The MIL and IVT methods perform poorly when the target object is occluded by a similar object (e.g., #154 and #190) because the adopted Harr-like features are less effective. The $l_1$ tracker also drifts away from the target after occluded by a similar object (e.g., #123 and #190). In contrast, our tracker performs stably (e.g., #256 and #382) in the entire sequence when there is a large scale change with heavy occlusion. Our tracker also does not drift away when the target reappears again (e.g., #123) because it is easy to differentiate the target and similar objects by using both holistic and local information. In the Occlusion2 sequence, our tracker is able to track the target accurately, especially when heavy occlusion (e.g., #180, #272 and #719) or in-plane rotation (e.g., #360 and #504) occurs. This attribute to the local histogram model has both spatial and partial information of the target object. FragTrack and $l_1$ methods can also perform well (e.g., #819) because partial occlusion is taken into account. The FragTrack method handles partial occlusion via the part-based representation with histograms, but it can not handles appearance change caused by pose and occlusion (e.g., #504). The $l_1$ tracker handles occlusion based on sparse representation with trivial templates. However, it can easily lead to tracking drift because the simple update method.

Figure 2a and b, respectively show the tracking results on sequences DavidIndoor and Singer1 with dramatic illumination changes. In the DavidIndoor sequence, the ambient light changes from dark to bright in the first few frames, the scale and pose of the object both change gradually (e.g., #60, #115 and #160). Our tracker and the IVT method can successfully track the target throughout the entire sequence (e.g., #60, #200 and #462). However, the Frag tracker drifts from the target (e.g., #115) due to the sudden and large illumination change and when the out-of-plane pose change happens, the $l_1$ method drifts away (e.g., #160 and #200) from the ground truth locations gradually. In the Singer1 sequence, a singer undergoes drastic appearance change due to illumination variation and scale change. The IVT, VTD and proposed methods perform well whereas the other

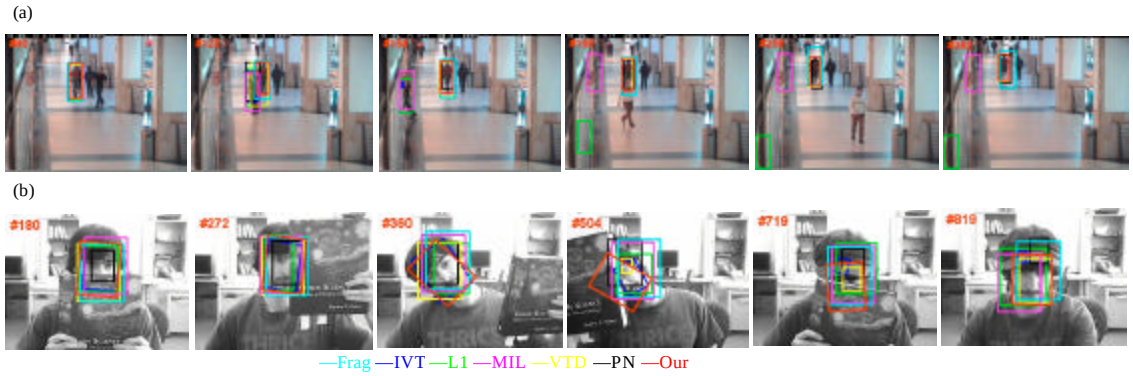


Fig. 1(a-b): Tracking results of 7 trackers on sequences, (a) Caviar and (b)Occlusion with severe occlusions delineated by different colors, Frame numbers are overlayed in red

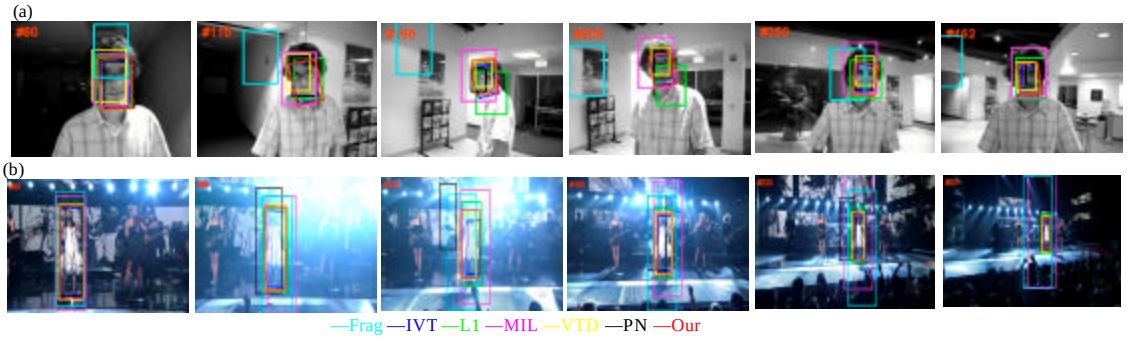


Fig. 2(a-b): Tracking results of 7 trackers on sequences, (a) DavidIndoor and (b) Singer1 with dramatic illumination change delineated by different colors. Frame numbers are overlayed in red

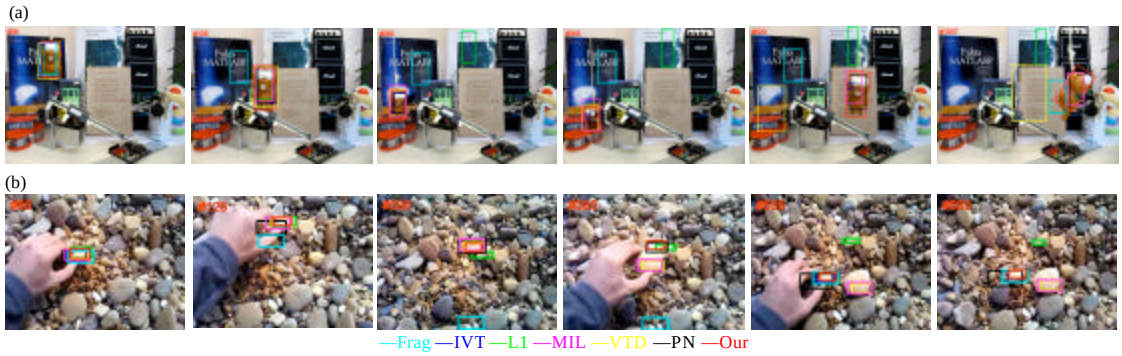


Fig. 3(a-b): Tracking results of 7 trackers on sequences, (a) Lemming and (b) Stone with background clutter delineated by different colors. Frame numbers are overlayed in red

methods drift away when drastic illumination change occurs (e.g., #86 and #125). The reason is that subspace learning method is robust to illumination changes. In this sequence, the PN method loses track of the target object for most of the frames (e.g., #125). We note that the $l_1$ tracker performs better than the IVT method, but also fails when the target experiences significant scale change and camera movement (e.g., #125). Moreover, we can find that the MIL and Frag methods do not estimate the scale change well (e.g., #160 and #233).

Figure 3a and b, respectively show the tracking results on sequences Lemming and Stone with background clutter. In the Lemming sequence, there are also partial occlusions, which add difficulty for visual tracking. From Fig. 3a, we observe that our tracker and the MIL method perform better than the other methods (e.g., #283 and #1037). The Frag, $l_1$ and VTD methods all lose track of the target gradually (e.g., #135, #283, #365 and #520). In addition, the IVT method fails when the target undergoes severe occlusion (e.g., #365) and the PN method is able to capture the target as long as there is no rotation or occlusion (e.g., #365 and #1037). In the Stone sequence, there are numerous stones of different shapes and colors. Most tracker fail as holistic representations inevitably include background pixels that may be considered as part of foreground object. The Frag, MIL and VTD methods drift away when the target object is occluded (e.g., #125, #393 and #555) whereas the IVT method and our tracker successfully track the location throughout the sequence (e.g., #300 and #593). The $l_1$ tracking method is very easy to drift away from the target in background clutters (e.g., #393 and #555). The PN method is able to recapture the target object again after drifting away but with higher tracking errors and lower success rate (e.g., #125).

Figure 4a and b, respectively show the tracking results on sequences Jumping and Deer with dramatic motion blur. Fast motion of the target object or the camera may lead to motion blur which is difficult to account for in object tracking. In the Jumping sequence, the target object is jumping and the motion blurs are very severe. From Fig. 4a, it can be seen that the proposed tracker performs better than other methods whereas the PN and MIL methods are also able to track the target object (e.g., #40, #150 and #304). We note that the PN method exploits the underlying structure of positive and negative samples to learning effective classifier which facilitates object tracking. Also the MIL method adopts multiple

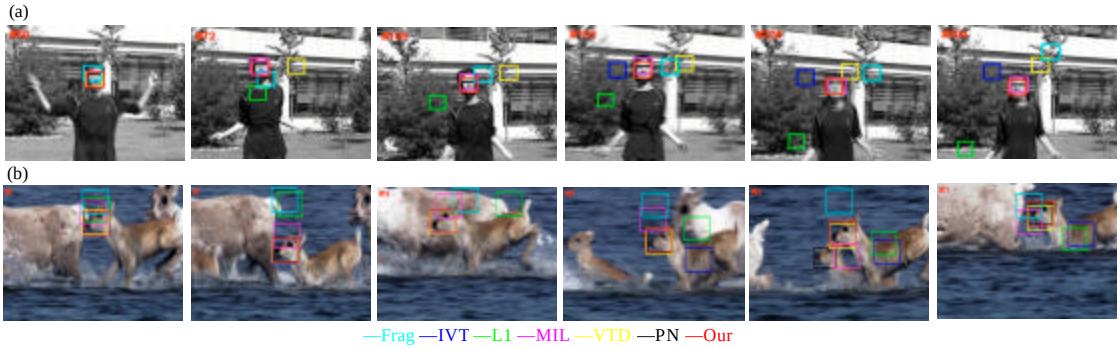


Fig. 4(a-b): Tracking results of 7 trackers on sequences, (a) Jumping and (b) Deer with motion blur delineated by different colors. Frame numbers are overlayed in red

Table 1: Success rates of 7 trackers on 9 different video sequences. On average, the proposed tracker outperforms the other 6 state-of-the-art trackers

| | Frag | IVT | L1 | MIL | VTD | PN | Our |
|---|---|---|---|---|---|---|---|
| Caviar1 | 0.68 | 0.28 | 0.28 | 0.25 | 0.83** | 0.70 | 0.91* |
| Occlusion2 | 0.60 | 0.59 | 0.67** | 0.61 | 0.59 | 0.49 | 0.85* |
| DavidIndoor | 0.20 | 0.71** | 0.63 | 0.45 | 0.53 | 0.60 | 0.81* |
| Singer1 | 0.34 | 0.66 | 0.70 | 0.34 | 0.79 | 0.41 | 0.86* |
| Lemming | 0.13 | 0.18 | 0.13 | 0.53** | 0.35 | 0.49 | 0.58* |
| Stone | 0.15 | 0.65** | 0.29 | 0.32 | 0.42 | 0.41 | 0.67* |
| Jumping | 0.14 | 0.28 | 0.09 | 0.53 | 0.08 | 0.69** | 0.70* |
| Deer | 0.08 | 0.22 | 0.04 | 0.21 | 0.58** | 0.41 | 0.69* |
| Girl | 0.69** | 0.43 | 0.33 | 0.52 | 0.51 | 0.58 | 0.70* |

*,** Show the best and second best results for each sequence, respectively

instance learning to develop a discriminative model which can deal with appearance change caused by motion blur. The IVT method is able to capture the target object in some frames (e.g., #72 and #100) but fails when there exits drastic image blur (e.g., #150 and #226). The Frag, $L_1$ and VTD methods have relative lager errors during tracking process. In the Deer sequence, we can see that most tracking methods fail to track the target object at the beginning of this sequence (e.g., #4 and #7), the Frag, MIL and $l_1$ methods fail when there is drastic motion blur (e.g., #19 and #37). In this sequence, the IVT method is also prone to drift when the motion blur occurs (e.g., #37 and #52). However, our tracker and the VTD method perform better than the other methods (e.g., #19, #37, #52 and #71). The tracking results on sequence Deer show that the robustness of our proposed method in fast motion and background clutters.

Figure 5 presents the tracking results of the evaluated trackers on sequence Girl. In this sequence, the challenging factors include in-plane rotation, out-of-plane rotation, 360 degree pose variation and partial occlusion by another face which is similar to the target. The IVT, VTD and PN tracking methods fail when the target object turns her head (e.g., #117 and #200) or is severely occluded (e.g., #322 and #433). As the holistic sparse representation method cannot deal with heavy occlusions and there is no drift alleviation mechanism, the $l_1$ tracking method does not perform well in this sequence (e.g., #82, #117, #322 and #433). Compared with other tracking method, the use of local histogram of our tracking method helps in accounting for appearance changes due to complex rotation. The experimental results show that our method is more robust and accurate.

## QUANTITATIVE EVALUATION

To quantitatively evaluation the robustness under challenging conditions, we measure the tracking success rate and center location error using the ground truth object locations obtained by manual labels at every 5 frames. We conduct quantitative comparisons between the proposed tracker and other algorithms using PASCAL VOC (Everingham *et al*., 2010) challenge criterion. Given the tracking result $R_T$ and the ground truth $R_G$, the score to evaluate the success rate is defined as:

$$\text{score} = \frac{\text{area}(R_T \cap R_G)}{\text{area}(R_T \cup R_G)} \quad (20)$$

and tracking in each frame is considered to be successful when the score is above 0.5. The success rates of our tracker and the other six algorithms on the challenging sequences are summarized in Table 1. For each sequence, the best and second best results are shown in red and blue respectively. From it, we can see that the proposed tracker outperforms the other trackers in all sequences.

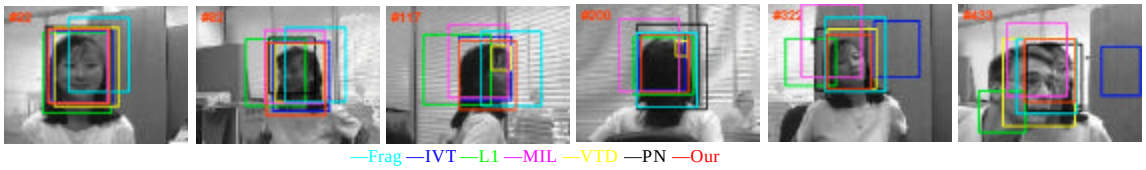


Fig. 5: Tracking results of 7 trackers on sequence Girl with rotation delineated by different colors. Frame numbers are overlayed in red

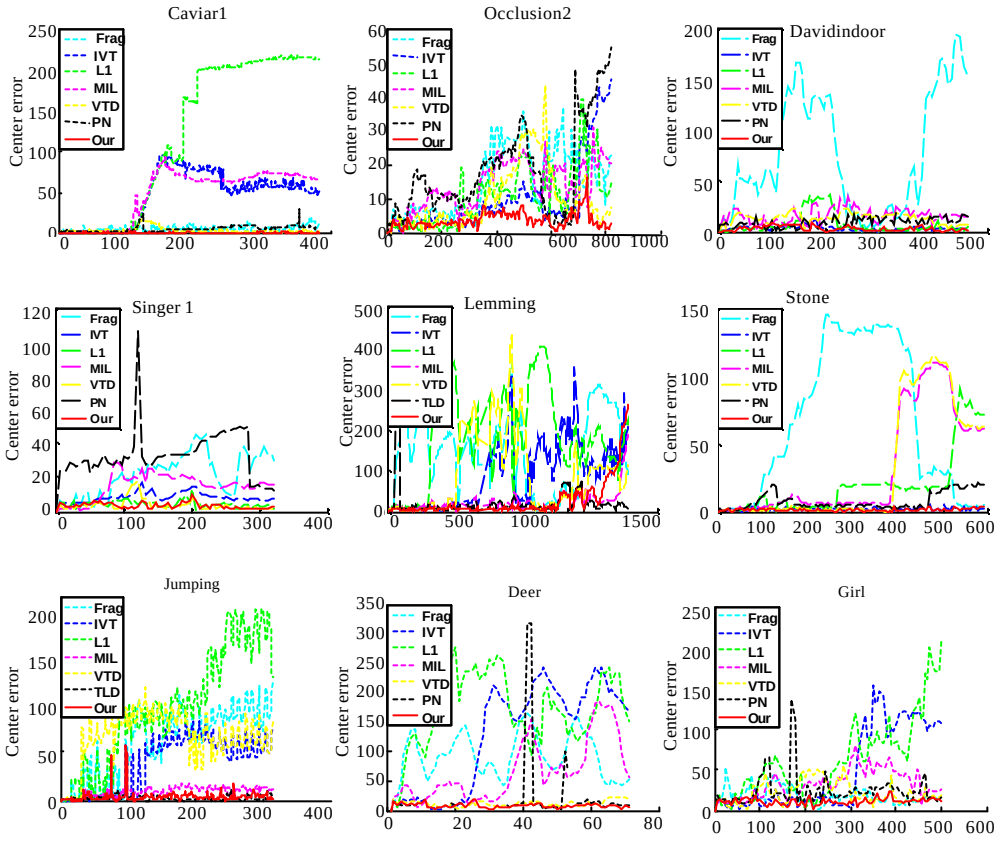


Fig. 6: Center location error between tracking result and ground truth over time for 7 trackers applied to 9 video sequences

The second criterion for evaluating the tracking performance is the center location error, which is based on the distance between the central of the tracking result and that of the ground truth. Figure 6 shows the center location errors of the evaluated tracking methods on the nine challenging sequences. From this *figure*, we can see that the proposed tracker consistently produce a smaller center location error than other tracking methods in general. This implies that the tracker can accurately track the target object despite severe occlusions, dramatic illumination change, background clutter, motion blur and rotation. The average center location errors are presented in Table 2. Similarly, the best and second best results are shown in red and blue, respectively. From these comparison results, we can see that our tracking method performs well against most of the evaluated trackers and has slightly higher tracking errors than the IVT method on Stone sequence and the PN method on Jumping and Lemming sequences. Overall, in thorough experiments involving nine challenging sequences and other six state-of-the-art trackers, the proposed algorithm demonstrates very promising tracking performance.

Table 2 : Average center location errors of 7 trackers on 9 different video sequences. On average, the proposed tracker outperforms the other 6 state-of-the-art trackers

| | Frag | IVT | L1 | MIL | VTD | PN | Our |
|---|---|---|---|---|---|---|---|
| Caviar1 | 5.7 | 45.2 | 119.9 | 48.5 | 3.9** | 5.6 | 0.9* |
| Occlusion2 | 15.5 | 10.2** | 11.1 | 14.1 | 10.4 | 18.6 | 3.2* |
| DavidIndoor | 76.7 | 3.6** | 7.6 | 16.1 | 13.6 | 9.7 | 3.4* |
| Singer1 | 22.0 | 8.5 | 4.6 | 15.2 | 4.1** | 32.7 | 3.7* |
| Lemming | 149.1 | 93.4 | 184.8 | 25.6** | 86.9 | 23.2* | 28.1 |
| Stone | 65.9 | 2.3* | 19.2 | 32.3 | 31.4 | 8.0 | 2.5** |
| Jumping | 58.4 | 36.8 | 92.4 | 9.9 | 63.0 | 3.6* | 4.9** |
| Deer | 92.1 | 127.5 | 171.5 | 66.5 | 11.9** | 25.7 | 8.1 |
| irl | 18.1** | 48.5 | 62.4 | 32.2 | 21.4 | 23.2 | 10.9* |

*,** Show the best and second best results for each sequence, respectively

## COLLUSION

In this study, we propose and demonstrate an effective and robust tracking in a co-training framework. Our algorithm can encode the holistic appearance model of the object in a compact linear subspace while strengthening the local histogram power. Moreover, the co-training framework helps the models update each other, which is especially helpful when each of them fails during tracking. Quantitative and qualitative evaluations demonstrate that our proposed tracking method performs well against the other state-of-the-art methods.

## ACKNOWLEDGMENTS

This work was supported by Chinese Forestry Industry Research Special Funds for Public Welfare (Grant No.201104090). The Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP)(20120161110014), New Century Excellent Talents in University (NCET-11-0134), National Natural Science Foundation of China (61072122), and Key Project of Hunan Provincial Natural Science Foundation (11JJ2053).

## REFERENCES

Adam, A., E. Rivlin and I. Shimshoni, 2006. Robust fragments-based tracking using the integral histogram. Proceedings of the International Conference Computer Vision and Pattern Recognition, June 17-22, 2006, IEEE., pp: 798-805.

Avidan, S., 2007. Ensemble tracking. IEEE Trans. Pattern Anal. Mach. Intell., 29: 261-271.

Babenko, B., M.H. Yang and S. Belongie, 2009. Visual tracking with online multiple instance learning. Proceedings of the International Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL., pp: 983-990.

Babenko, B., M.H. Yang and S. Belongie, 2011. Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell., 33: 1619-1632.

Black, M.J. and A.D. Jepson, 1998. Eigen tracking: Robust matching and tracking of articulated objects using a view-based representation. Int. J. Comput. Vision, 26: 63-84.

Collins, R.T., Y. Liu and M. Leordeanu, 2005. Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Machine Intell., 27: 1631-1643.

Everingham, M., L. Gool, C. Williams, J. Winn and A. Zis-Serman, 2010. The pascal Visual Object Classes (VOC) challenge. Int. J. Comput. Vis., 88: 303-338.

Grabner, H., C. Leistner and H. Bischof, 2008. Semi-supervised on-line boosting for robust tracking. Proceedings of the 10th European Conference on Computer Vision, October 12-18, 2008, Marseille, France, pp: 234-247.

Grabner, H., M. Grabner and H. Bischof, 2006. Real time tracking via online boosting. Proceedings of the British Machine Vision Conference, September 5, 2006, Edinburgh, pp: 47-56.

Kalal, Z., J. Matas and K. Mikolajczyk, 2010. P-n learning: bootstrapping binary classifier by structural constraints. Proceedings of the International Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA., USA., pp: 49-56.

Kwon, J. and K.M. Lee, 2010. Visual tracking decomposition. Proceedings of the Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA., USA., pp: 1269-1276.

Li, Y., H. Ai, T. Yamashita, S. Lao and M. Kawade, 2008. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans. IEEE Trans. Pattern Anal. Mach. Intell., 30: 1728-1740.

Matthews, L., T. Ishikawa and S. Baker, 2004. The template update problem. IEEE Trans. Pattern Anal. Mach. Intell., 26: 810-815.

Mei, X. and H. Ling, 2009. Robust visual tracking using $L_1$ minimization. Proceedings of the 12th International Conference on Computer Vision, September 29-October 2, 2009, Kyoto, pp: 1436-1443.

Ross, D.A., J. Lim, R.S. Lin and M.H. Yang, 2008. Incremental learning for robust visual tracking. IJCV, 7: 125-141.

Wang, S., H. Lu, F. Yang and M.H. Yang, 2011. Superpixel tracking. Proceedings of the International Conference on Computer Vision, November 6-13, 2011, Barcelona, pp: 1436-1443.

Zhang, H. and L. Liu, 2013. Recovering low-rank and sparse components of matrices for object detection. Electron. Lett., 49: 109-111.