Small target detection in infrared video sequence using robust dictionary learning

Chunwei Yang\textsuperscript{a,b,c}, Huaping Liu\textsuperscript{b,c,*}, Shouyi Liao\textsuperscript{a}, Shicheng Wang\textsuperscript{a}

\textsuperscript{a} High-Tech Institute of Xi’an, Xi’an, Shaanxi 710025, PR China\textsuperscript{b} Department of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China\textsuperscript{c} State Key Lab. of Intelligent Technology and Systems, TNLIST, Beijing 100084, PR China

\textbf{ABSTRACT}

Small target detection in infrared video sequence is a challenging problem. In this paper, a collaborative structured sparse coding (SSC) model which incorporates the $L_1, L_2$ and $L_2, 1$ regularization terms is proposed. The Alternating Direction Method of Multiplier (ADMM) is developed to solve this model. Further, online dictionary learning is embedded into the model and temporal information is incorporated to eliminate the clutters and noises. Extensive synthetic and real data experiments show that our method obtains better detection performance than baseline methods and state-of-art infrared-patch-image (IPI) model.

\textbf{1. Introduction}

Infrared imaging technology has the advantages of large dynamic range, long operating distance and all-day use and it is widely used in many areas such as space surveillance systems, early-warning systems, and missile tracking systems. Infrared small target detection technology plays an important role in these applications and has developed rapidly in recent years. However, small target just occupies very few pixels in the infrared image and when there exists a complex background such as sea clutters and cloud clutters, small target is usually buried with low signal-to-clutter ratio (SCR). What’s more, there’s no information about concrete shape and texture because of long imaging distance. All of these make small target detection in infrared images a rather difficult and challenging problem [1–4].

In general, small target detection methods are divided into two classes: the single frame detection methods and the sequential detection methods. Some typical recently proposed methods are listed as follows. Refs. [2,5] formulated the infrared small target detection as salient region detection and presented a robust directional saliency-based method using phase spectrum of Quaternion Fourier Transform. Refs. [6,7] showed a plausible computational model integrating the robust properties of human visual system (HVS) using Laplacian scale space theory and Tune–Max based optimization method. Ref. [8] proposed an improved algorithm based on the contrast mechanism of HVS, which exploited Laplacian of Gaussian (LoG) filter to deal with input image and processed the filtered image with morphological method in all directions. Ref. [9] presented an effective small target detection algorithm inspired by the contrast mechanism of HVS and derived kernel model, which used local contrast measure (LCM) to obtain the local contrast map of the input image. Ref. [10] proposed a novel method which used Difference of Gaussians (DoG), Gaussian window and Proportional–Integral–Derivative (PID) algorithm combining contrast mechanism, visual attention and eye movement mechanism of HVS.

On the other hand, sparse coding strategy [11–15] which is a hot topic in recent years, has been widely used in image denoising,
image feature extraction and pattern recognition, etc. Ref. [3] proposed an infrared patch-image (IPI) model and formulated target detection as an optimization problem of recovering low-rank and sparse matrices, which can be effectively solved using stable principal component analysis. Ref. [16] presented a sparse-representation-based automatic target detection approach for forward-looking infrared (FLIR) imagery, in which the sparsity is calculated through approximately expressing the pixels within a center image block as the sparsest representation of its surroundings in feature domain. However, they are aimed at solving the target detection problem in a single frame.

In this paper, we develop a method called robust dictionary learning (RDL) for small target detection in infrared video sequence. The main contributions of this paper are summarized as follows:

(1) A collaborative convex structured sparse coding (SSC) model is proposed to address the infrared small target detection problem. In this model, a $L_{1,2}$ penalty term is used to discover the locations of the small target and $L_{2,1}$ penalty term is used to characterize the background.

(2) An iterative optimization method which is based on Alternating Direction Method of Multiplier (ADMM) is developed to solve the proposed model.

(3) Online dictionary learning and temporal information are incorporated to eliminate clutters and noises.

The remainder of this paper is organized in the following way: In Section 2, we give the overview of the RDL methodology. In Section 3, we show the role of RDL strategy in small target detection in video sequence. In Section 4, we develop the optimization algorithms to solve the SSC model. In Section 5, we give the experimental results. Finally, the conclusions and perspectives are given in Section 6.

Notations: We use various matrix norms. Here are the notations we use: $\|\mathbf{M}\|_F$ is the Frobenious norm, which is also equal to $\sqrt{\text{Tr}(\mathbf{M}^T \mathbf{M})}$; $\|\mathbf{M}\|_{2,1}$ is the sum of the $L_2$ norm of the rows of $\mathbf{M}$; and $\|\mathbf{M}\|_{2,2}$ is the sum of the $L_2$ norm of the columns of $\mathbf{M}$.

2. Overview of the proposed method

Fig. 1 shows the overview of the RDL method for small target detection in infrared video sequence. Firstly, a window is used to slide from left and top to right and down in each image frame of the video sequence to obtain a series of image blocks. Obviously the size and the number of the block depends on the size and the sliding step of the sliding window respectively. We vectorize each block as a column of a new matrix as $\mathbf{d}_{1}, \mathbf{d}_{2}, \ldots, \mathbf{d}_{N}$, where $N$ is the number of blocks. Then we decompose $\mathbf{D} = [\mathbf{d}_{1}, \mathbf{d}_{2}, \ldots, \mathbf{d}_{N}]$ using SSC strategy, after which we can obtain saliency which denotes the information of the target and noises plus the exemplars which represent the background. After that, we introduce online dictionary learning which means that the representatives selected from the exemplars obtained from the $t$-th frame are embedded into the dictionary of the ($t+1$)-th frame. By this way, the dictionary is updated online in the whole video sequence. To this end, in order to eliminate noises, the temporal information is exploited inspired by the fact that the target always has a relatively standard trajectory, while the noises appear in random locations (illustrated in Fig. 2). Finally, we can obtain the small target from the infrared video sequence.

3. Target detection using structured sparse coding

In this section, we introduce collaborative sparsity to small target detection in which background and target(or noises) are transformed to row sparsity and column sparsity of matrices, respectively.

3.1. Small target detection in a single frame

As to the small target detection, the infrared image can be divided into two parts: background and target. Considering the matrix $\mathbf{D}_t = [\mathbf{d}_{1}, \mathbf{d}_{2}, \ldots, \mathbf{d}_{N}] \in \mathbb{R}^{n \times N}$ which is obtained through partitioning the $t$-th frame of the video sequence, where each column vector denotes a vectorized block. We find an optimal subset $\mathbf{D}_t = [\mathbf{d}_{1}, \mathbf{d}_{2}, \ldots, \mathbf{d}_{M}] \in \mathbb{R}^{m \times M}$ ($M < N$) to approximately reconstruct the matrix $\mathbf{D}_t$ and the value of $M$ is expected to be as small as possible. Therefore, the subset should be representatives of all the samples, i.e., the cost of using such subset to reconstruct the whole data set should be small. That is to say, for any $k \in [1, N]$, the sample $\mathbf{d}_{(k)}$ should be well approximated by

$$\mathbf{d}_{(k)} = \sum_{i=1}^{m} \omega_{k,i} \mathbf{d}_{(i)} + \sum_{i=M+1}^{N} \omega_{k,|M+1|} \mathbf{d}_{(i)}$$

(1)

where $\omega_{k,1}, \ldots, \omega_{k,|M|}$ are the reconstruction coefficients for $\mathbf{d}_{(k)}$.

To characterize the reconstruction capability, the following cost should be minimized to use such subset to reconstruct the whole sample set:

$$\min_{\mathbf{D}_t} \|\mathbf{D}_t - \mathbf{D}_t \mathbf{W}_t\|_F^2$$

(2)

where $\mathbf{W}_t \in \mathbb{R}^{m \times M}$ is the coefficient matrix of the $t$-th frame.

![Fig. 1](image_url) The overview of RDL method in this paper. The infrared video sequence is firstly transformed into blocks which are vectorized as a dictionary. Then we obtain exemplars representing background and saliency denoting the target and noises by SSC. After that, representatives are embedded into the next dictionary to achieve the online learning. Meanwhile temporal information is incorporated to eliminate the noises and the location of the real target can be obtained.
It is obvious that we want few atoms to participate in the approximation, but we want each atom to contribute to as many columns of the sample matrix as possible. In other words, most rows of the coefficient matrix should be zero, but the nonzero rows should have many nonzero entries. Therefore, a straightforward approach is to minimize the following objective function,

$$\min_{W_t} \|W_t\|_{\text{row-0}} + \lambda \|D_t - D_t W_t\|_F^2$$

where $W_t \in \mathbb{R}^{N \times N}$ is the pursuit coefficient matrix; the term $\|D_t - D_t W_t\|_F^2$ is used to evaluate the reconstruction error, and the parameter $\lambda$ is used to balance different penalty terms. The symbol $\|W_t\|_{\text{row-0}}$ counts the number of nonzero rows of $W_t$. By adding such a term into the objective function, the trivial solution can be avoided and the obtained solution will be row sparse, i.e., most of its rows are zero vectors.

In the model (3), the sparsity is imposed on the rows of the matrix $W_t$, while the reconstruction error is evaluated using the Frobenius norm.

To isolate the target, we exploit the fact that the target should not be well reconstructed by the subset and therefore the corresponding reconstruction error should be large. That is to say, if the target is located in the $t$-th sample, then the $t$-th column of the reconstruction error matrix $D_t - D_t W_t$ will be nonzero and may admit large values. Otherwise, if the $t$-th sample is background, then the error matrix $D_t - D_t W_t$ will be close to zero. On the other hand, it is frequently observed that the targets are usually a minority in the sample set, i.e., the number of the targets is usually not large. We make use of this property and formalize it as the column sparsity regularization term

$$\|D_t - D_t W_t\|_{\text{column-0}}$$

which counts the number of the nonzero column of $D_t - D_t W_t$.

Therefore, we modify the model (3) as:

$$\min_{W_t, E_t} \|W_t\|_{\text{row-0}} + \lambda \|E_t\|_{\text{column-0}} \quad \text{s.t. } D_t - D_t W_t = E_t$$

In such a model, the row-sparsity is imposed on $W_t$ to detect the exemplars and the column sparsity is imposed on the error matrix $E_t$ to isolate the target. By this way, the extracted exemplars will focus on reconstructing the background only. This method is illustrated in Fig. 3. Since the two sparsity optimization problems are simultaneously solved, we call it as collaborative sparsity.

However, such an approach is of little practice use, since the optimization problem is NP-hard as its solution requires a combinatorial search which grows faster than polynomial as the dimension $N$ grows. An alternative is to use the $L_{2,1}$-norm and $L_{1,2}$-norm to approximate the $\|\cdot\|_{\text{row-0}}$ and $\|\cdot\|_{\text{column-0}}$, respectively [17–19]. This results in the following convex optimization problem:

$$\min_{W_t, E_t} \|W_t\|_{2,1} + \lambda \|E_t\|_{1,2} \quad \text{s.t. } D_t - D_t W_t = E_t$$

The above optimization problem can be efficiently solved [20]. After obtaining the value of $E_t$, the $L_2$ norm of the $j$-th column for $E_t$, which is denoted as $\|E_t\|_{1,2}$, is used to evaluate the possibility of the $j$-th sample to be the saliency of the $t$-th frame and the weight $\|W_t\|_2$ can be reliably utilized to extract exemplars. After deriving $W_t$, we use the weight $\|W_t\|_2$ to rank the samples. The larger $\|W_t\|_2$, the more important this sample is.

In such a collaborative SSC model, as to the $t$-th frame, the column sparsity is imposed on the matrix $E_t$ to isolate the saliency and the row sparsity is imposed on $W_t$ to detect the exemplars. By this way, the saliency is obviously obtained and the extracted exemplars focus on reconstructing the background.

### 3.2. Robust dictionary learning

An important problem in the detection process is to update the dictionary when the new frame is coming. In order to find the representatives for the new frame, we let $D_{t+1} = [S, D_{t+1}]$, where $S$ is a representative exemplars selected from the exemplars of the $t$-th frame in video sequence that we have already found (the samples corresponding to the $k$ largest values of Fig. 4(b)). Now, we use $D_{t+1}$ instead of $D_t$, so that they can well describe the coming frame. In this way, the dictionary can be updated online in the whole process.

Therefore, model (6) can be modified as:

$$\min_{W_{t+1}, E_{t+1}} \|W_{t+1}\|_{2,1} + \lambda \|E_{t+1}\|_{1,2}, \quad \text{s.t. } D_{t+1} - D_{t+1} W_{t+1} = E_{t+1}$$

### 3.3. Temporal information

The two largest values of Fig. 4(b) represent the locations of saliency that corresponds to either target or noise. In order to distinguish them, we make use of the trajectory information which is
are dual variables (i.e., the Lagrangian multipliers), and \( l + Y_l \) is the distance to the target location in the former frame \( DW \).

To embed the temporal information into the model, we extend Eq. (7) as the following:

\[
\min_{W_{t+1},E_{t+1}} \|W_{t+1}\|_{2,1} + \lambda \|P_{t+1}E_{t+1}\|_{1,2}, \quad \text{s.t.} \quad D_{t+1} - D_{t+1}W_{t+1} = E_{t+1}
\]

where \( P_{t+1} \) is a diagonal matrix of the \((t+1)\)-frame, whose diagonal element is \( \rho_{ij} \). The weight \( \rho_{ij} \) reflects the temporal characteristic of the video sequence. In this work we use this strategy to effectively eliminate noises, which is illustrated in Fig. 2). Then the real small target can be obtained.

4. Optimization algorithm

The optimization problem in (8) is intrinsically convex and therefore we adopt the ADMM [20] to solve Eq. (8). To this end, we transform the above optimization problem as

\[
\min_{G,W} \|G\|_{2,1} + \lambda \|PE\|_{1,2}, \quad \text{s.t.} \quad D - DW = E, W = G
\]

The augmented Lagrangian associated with the above optimization problem is given by

\[
L(G,W,E,Y_1,Y_2) = \|G\|_{2,1} + \lambda \|PE\|_{1,2} + \text{Tr}(Y_1^T(D - DW - E)) + \frac{\mu}{2} \|D - DW - E\|_F^2 + \text{Tr}(Y_2^T(W - G)) + \frac{\mu}{2} \|W - G\|_F^2
\]

where \( Y_1 \) and \( Y_2 \) are dual variables (i.e., the Lagrangian multipliers), \( \mu \) is a positive scalar. In order to find a minimizer of the constrained problem (10), the ADMM algorithm uses a sequential iterations (here, the superscripts \((k+1)\) and \(k\) mean that the variables obtained from the \((k+1)\)-th and \(k\)-th iteration, respectively)

\[
\begin{align*}
G^{k+1} & = \arg \min_G L(G,W^{k+1},E^{k+1},Y_1^{k+1},Y_2^{k+1}) \\
W^{k+1} & = \arg \min_W L(G^{k+1},W,E^{k+1},Y_1^{k+1},Y_2^{k+1}) \\
E^{k+1} & = \arg \min_E L(G^{k+1},W^{k+1},E,Y_1^{k+1},Y_2^{k+1}) \\
Y_1^{k+1} & = Y_1^k + \mu(D - DW^{k+1} - E^{k+1}) \\
Y_2^{k+1} & = Y_2^k + \mu(W^{k+1} - G^{k+1})
\end{align*}
\]

until \( \|D - DW^{k+1} - E^{k+1}\|_F \leq \varepsilon \) and \( \|W^{k+1} - G^{k+1}\|_F \leq \varepsilon \), where \( \varepsilon \) is the tolerance error. In the following we explain how to solve the optimization problems in (12).

First, the optimization over \( G \) is equivalent to

\[
\min_G \frac{\lambda}{\mu} \|G\|_{2,1} + \|G - V\|_F^2
\]

where \( V = \frac{1}{\mu} Y_2^{k+1} + W^{k+1} \). According to [19], the \( i \)-th row of the optimal solution \( G \) can be analytically obtained as

**Algorithm 1.**

**Initialize** Data set \( D \in \mathbb{R}^{d \times N} \)

**Ensure:** Solutions \( W \in \mathbb{R}^{d \times N} \) and \( E \in \mathbb{R}^{d \times N} \)

1. Initialization: Set \( G, W, E, Y_1 \) and \( Y_2 \) as zero matrices with appropriate dimensions.
2. **while** Not convergent **do**
3. Update \( G, W, E, Y_1 \) and \( Y_2 \) according to (12).
4. **end while**

\[
G^{(i)} = \begin{cases} 
(1 - \frac{1}{\mu} \|V\|_1) V & \text{when } \|V\|_2 > \frac{1}{\mu} \\
0 & \text{otherwise}
\end{cases}
\]
5. Experimental results and analysis

To evaluate our proposed method, we use the following five baseline methods for comparison: (I) Tophat filtering method [21], (II) MaxMedian filtering method [22], (III) MaxMean filtering method [22], (IV) IPI model [3] and (V) frame difference algorithm, the first four regard to the case of single frame and the last aims at dealing with moving target detection. By such a comparison we can clearly show the superior performance of RDL.

5.1. Simulation datasets

In order to verify the proposed approach in this paper, four simulation databases are constructed by using 40 real infrared background images and four targets generated artificially. The background images are chosen from a real infrared video sequence. The video sequence in each database includes 40 infrared images. The targets in database 1 and database 2 are in sky and sea, respectively. However, except for adding noise which is the same as target generated in random location, dataset 3 and database 4 are the same as database 1 and 2, respectively.

A synthesized image \( I \) with target and noise can be achieved by embedding a target image \( T \) and a large noise image \( N \) with size of \( m \times n \) and \( p \times q \) respectively into a background image \( B \). While the image with target only is constructed in the same way, only to delete the noise synthesized process. Take the former for example. The detail is as follows:

\[
I(x, y) = \begin{cases} 
\max(T(x-x_0,y-y_0),B(x,y)) & \text{when } x \in (1+x_0,m+x_0), \\
\max(N(x-x',y-y'),B(x,y)) & \text{when } x \in (1+x',p+x'), \\
B(x,y) & \text{otherwise}
\end{cases}
\]

where \( (x_0,y_0) \) and \( (x',y') \) is the trajectory of the target and a randomly produced pixel location of the noise, respectively, which the left upper corner of the image \( I \) corresponds to in the image \( B \).
and N. Then we blur the synthesized image using Gaussian filter to make it close to a real one. Finally, we obtain database 1 and 2 without large noises and database 3 and 4 with large noises. In the databases, the size of target is different in order to test the robustness of the method, i.e. $3 \times 3$, $6 \times 6$, $2 \times 6$, $3 \times 9$. Fig. 5 shows the examples of the synthesized images.

5.2. Qualitative evaluation

In this section, we list the infrared small target detection results of simulation databases. In addition, a public infrared sequence is tested by the above methods. The results are shown in Figs. 6 and 7.

From Fig. 6, we can see that all the methods can detect the target, but there exists too many clutters in the results of Tophat, MaxMedian, MaxMean filtering methods and frame difference algorithm, which shows the superiority of IPI and RDL, but RDL is better than IPI to some extent. The reason is that sea waves are similar to small target, and Tophat, MaxMedian and MaxMean filtering methods cannot eliminate the sea waves completely. While the sea waves are changing, frame difference algorithm cannot work well, too. When there exists a false target which is shown in database 3 and 4, all the baseline methods cannot identify the real target except RDL. The reason is that RDL successfully exploits the intrinsic temporal relation and therefore provides better
robustness. As to Fig. 7, when the background includes heavy clutters, all the baseline methods fail to suppress the clutters completely.

5.3. Quantitative evaluation

For comparison, a common evaluation indicator is adopted in this section, i.e. SCR Gain as follows [22,23,24]:

\[
\text{SCR Gain} = \frac{\text{SCR}_{\text{out}}}{\text{SCR}_{\text{in}}}
\]

(20)

here SCR is defined as follows:

\[
\text{SCR} = \frac{\mu_t - \mu_b}{\sigma_s}
\]

(21)

where \(\mu_t\) is the average value of the target, \(\mu_b\) and \(\sigma_s\) are the average value and the standard deviation of the pixels of the

---

**Table 1**

<table>
<thead>
<tr>
<th>Methods</th>
<th>SCR Gain of database 1</th>
<th>SCR Gain of database 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tophat</td>
<td>19.0204</td>
<td>22.1345</td>
</tr>
<tr>
<td>MaxMedian</td>
<td>19.0485</td>
<td>13.1699</td>
</tr>
<tr>
<td>MaxMean</td>
<td>2.2012</td>
<td>1.6343</td>
</tr>
<tr>
<td>IPI</td>
<td>320.4436</td>
<td>301.5273</td>
</tr>
<tr>
<td>Frame difference</td>
<td>3.2015</td>
<td>5.3314</td>
</tr>
<tr>
<td>RDL</td>
<td>370.8105</td>
<td>356.2737</td>
</tr>
</tbody>
</table>

Fig. 7. The results of infrared sequences: (a) original frames, detection results of (b) Tophat, (c) Maxmedian, (d) MaxMean, (e) IPI, (f) frame difference, and (g) RDL.
background, respectively. In general, the higher the SCR Gain is, the easier the infrared small target can be detected, which means that the processed method is better.

Table 1 gives the quantitative evaluation results of the five baseline methods and RDL method. From Table 1, we can obviously see that RDL method is better than the compared ones. The main reason is that the baseline methods cannot suppress clutters effectively, leading to large $\sigma_r$ and small SCR.

Furthermore, in Fig. 8 we show receiver operating characteristic (ROC) curves of each method for database 2 and database 4 to provide a quantitative comparison of detection performance. The ROC curves describe the probability of detection (PD) as a function of the probability of false alarms (PFA), which are more specific in describing the performance of target extraction using a threshold on processed high SCR maps and more significant than SCR Gain. PD and PFA are defined as:

$$PD = \frac{\text{Number of real targets detected}}{\text{Number of real targets}} \quad (22)$$

$$PFA = \frac{\text{Number of false targets detected}}{\text{Total number of frames of video}} \quad (23)$$

For database 2, the temporal information is not used in which case RDL is applied to single frame. From Fig. 8(a), we can see that RDL and IPI get better performance than the other four baseline methods. For database 4 in which there exists a noise in each frame, as is illustrated in Fig. 8(b), RDL can distinguish the real target and outperforms the baseline methods to a large extent.

6. Conclusions and perspectives

In this paper, a novel method called robust dictionary learning is proposed to solve the small target detection in infrared video sequence. The $L_{1,2}$ and $L_{2,1}$ penalty term are used to discover the location of the small target and characterize the background, respectively, and the online dictionary learning and temporal information are used to eliminate the clutters and noises. Experiments show that RDL method outperforms the baseline methods, e.g., Tophat, MaxMedian, MaxMean filtering methods, IPI model and frame difference algorithm.

However, there are two limitations of our method. The one is that when the noise is close to the target, our method may fail to identify the real target. The other is that we only focus on single target case. How to solve these problems is our future work.

**Conflict of interest**

There is no conflict of interest.

**Acknowledgement**

This work was supported by the National Key Project for Basic Research of China under Grant 2013CB329403, and National Natural Science Foundation of China under Grant 61401470.

**References**


