Object detection and tracking algorithms using brain-inspired model and deep neural networks

Y Song*1,2, YF Zhao1,2, X Yang1,2, Y Zhou1,2, F N Wang1,2, Z S Zhang1,2, and Z K Guo1,2

1School of Optics and Photonics, Beijing Institute of Technology, Beijing, 100081 China;
2Beijing Key Laboratory for Precision Optoelectronic Measurement Instrument and Technology, Beijing, 100081 China

yongsong@bit.edu.cn

Abstract. As the most effective bio-intelligence system, Human Visual System (HVS) has significant advantages in image processing, which helps to solve the problems in infrared target detection and tracking, such as dim small target, complex background, target occlusion and appearance changes, etc. In this paper, several brain-inspired models (including lateral inhibition, receptive field, synchronous burst, visual attention, and cognitive memory) and Deep Neural Networks (DNNs) have been studied, and the corresponding algorithms are proposed, which include: an infrared target detection algorithm based on lateral inhibition and singular value decomposition, an infrared target detection algorithm based on receptive field and lateral inhibition, an infrared moving dim target detection algorithm based on ALI-PCNN, an infrared target detection algorithm based on GCF-SB visual attention model, a kernel correlation filtering target tracking algorithm based on multi-channel memory model, and a robust and efficient discriminative-correlation-filter-based tracking approach based on the Response Map Analysis Network. Our experimental results show that the proposed algorithms are beneficial to achieve accurate infrared target detection and robust tracking under complex conditions.

1. Introduction

Recent IR object detection and tracking algorithms face several problems. For IR object detection, one problem is that the captured infrared target generally has a complex background. Another is that the detection methods are not suitable for various size. At present, the existing methods generally detect infrared target from perspectives of spatial domain [1], frequency domain [2], mathematical morphology [3] and two-dimensional least mean-square (TDLMS) [4]. However, these methods have poor detection ability when detecting IR target from complex background. And it is hard to find a detection method which is appropriate for both small and area target in the existing infrared target detection methods. For the IR target tracking, it mainly meets two main challenges, which include: (1) the challenges caused by the environment, such as the serious occlusions, illumination variations, background clutters. (2) The challenges caused by the target itself, such as the target geometric deformation, rotation and gesture variations [5].
Applying the signal processing mechanisms of Human Visual System (HVS) in IR object detection and tracking field is beneficial to improve the anti-interference ability as well as the detection and tracking ability. Among lots of the HVS models, Lateral Inhibition (LI), Receptive Field (RF) and Cognitive Memory are of significance. The application of LI, RF and Cognitive Memory will do great helpful to detect and tracking target of various size in complex background.

In this paper, to detect and track infrared target in complex background, several algorithms are proposed, which include: an infrared target detection algorithm based on lateral inhibition and singular value decomposition, an infrared target detection algorithm based on receptive field and lateral inhibition, an infrared moving dim target detection algorithm based on ALI-PCNN, an infrared target detection algorithm based on GCF-SB visual attention model, a kernel correlation filtering target tracking algorithm based on multi-channel memory model, and a robust and efficient discriminative-correlation-filter-based tracking approach based on the Response Map Analysis Network. Experimental results show that applying the brain-inspired models and DNNs to the infrared target detection and tracking is beneficial to achieve the accurate infrared target detection and robust tracking under complex conditions.

2. Algorithms based on brain-inspired models

2.1. IR object detection algorithm based on lateral inhibition and singular value decomposition

2.1.1. Lateral inhibition network A 2D LI network is

\[
G(x, y) = F(x, y) - \sum_{m=-l}^{l} \sum_{n=-l}^{l} h(m, n) F(x + m, y + n)
\]

where \(F(x, y)\) and \(G(x, y)\) are the input and output, respectively, \(h(m, n)\) is LI coefficient, and \(l\) denotes inhibitory field radius.

An anisotropic Gauss kernel function was adopted to determine LI coefficients. If the scales in \(x\) and \(y\) directions are different, the \(x-y\) plane of Gauss kernel function will be an ellipse, as

\[
G(x, y, \sigma) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left[ -\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) \right],
\]

where \(\sigma_x^2\) and \(\sigma_y^2\) are the variances of \(x\) and \(y\) directions, respectively. Rotate the ellipse by \(\theta\), the \(x-y\) is transformed into \(u-v\) by

\[
\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},
\]

where \(\theta\) is the rotation angle of anisotropic Gaussian filter. Combine above equations, the LI coefficient can be calculated by

\[
G_\theta(u, v; \sigma_u, \sigma_v, \theta) = \frac{1}{2\pi\sigma_u\sigma_v} \exp \left[ -\frac{1}{2} \left( \frac{1}{\sigma_u^2} \left( x \cos \theta + y \sin \theta \right)^2 + \frac{1}{\sigma_v^2} (-x \sin \theta + y \cos \theta)^2 \right) \right]
\]

The SVD of the local gradient field is calculated to determine the direction parameters in LI. First, obtain gradient of \(I(x, y)\) at \((x_i, y_i)\) by

\[
\nabla I(i) = \nabla I(x_i, y_i) = \left[ \frac{\partial I(x_i, y_i)}{\partial x}, \frac{\partial I(x_i, y_i)}{\partial y} \right]^T,
\]

Then, sort \(\nabla I(i)\) into an \(n \times 2\) matrix \(G\) and perform SVD operations on it by
where \( n \) is the amount of pixels in each block, \( U \) is an orthogonal matrix with the size of \( n \times n \), \( S \) is a matrix with the size of \( n \times 2 \) and \( \lambda_1 \) and \( \lambda_2 \) reflect the energy of the eigenvector direction. \( V \) is a matrix with the size of \( 2 \times 2 \). The dominant orientation \( O_d \) can be achieved by

\[
O_d = \arctan \left( \frac{V_1}{V_2} \right).
\]

2.1.2. Local structure descriptor There are generally three situations in an image: (1) in the plat area, \( \lambda_1=\lambda_2\approx 0 \); (2) in the edge area, \( \lambda_1>\lambda_2\approx 0 \); (3) in the detailed area, \( \lambda_1>\lambda_2>0 \). In the proposed method, to classify the pixels, an LSD \( D_{LS} \) \((0 \leq D_{LS} \leq 1)\) is established. The larger \( D_{LS} \) is, the greater changes of the image.

\[
\begin{aligned}
D &= \lambda_1 + \lambda_2 \\
D_{LS} &= \frac{D - D_{min}}{D_{max} - D_{min}}.
\end{aligned}
\]

2.1.3. Modified adaptive LI network In the proposed method, \( O_d \) is adopted to determine the direction parameter \( \theta \) in Eq. (4). In the case that \( O_d \) does not exist, set \( \theta=0 \) to suppress clutters. In additional, set \( \sigma_u=C_1=1 \) and \( \sigma_v=C_2=1 \) according to [7], [8]. Finally, we can obtain the LI coefficients by

\[
G_y(\theta) = \frac{1}{2\pi C_1 C_2} \exp \left[ -\frac{1}{2} \left( \frac{1}{C_1^2} (x \cos O_d + y \sin O_d)^2 + \frac{1}{C_2^2} (-x \sin O_d + y \cos O_d)^2 \right) \right].
\]

Then a modified LI network can be achieved by the processes shown in figure 1, in which \( F(x, y) \) is input image (a) and the corresponding 3D plot (b), \( D_{LS}(x, y) \) in Figure 1(c) denotes the LSD, \( R(x, y) \) in figure 1(d) is the result of adaptive LI, and \( G(x, y) \) in Figure 1(e) is the final result.

**Figure 1.** The modified adaptive LI.
2.1.4. Algorithm process Because noises may affect the dominant orientation [9], we adopted a filter process as pre-process step. The process are presented in figure 2. Initially, the local block can be decomposed by SVD, and $S$ and $V$ are obtained. Then, the $D_{LS}$ (LSD) can be obtained according to $S$. Next, $\theta$ (rotation angle) can be determined by $O_d$ (dominant orientation). Moreover, to make the target more obvious, gray value compensation of the image by

$$K = \frac{255 \cdot n}{\sum_{i=1}^{n} G_{order}(i)}, i = 1, 2, \ldots, n$$

where $n$ is the amount of pixels with higher gray value in $G_{order}$, $K$ is the compensation coefficient.

**Figure 2.** The process of the proposed method.

2.2. IR object detection algorithm based on receptive field and lateral inhibition

2.2.1. Receptive field The two-dimensional Gabor function [10] can be used for describing the model of RF, we take its real component to model the spatial properties of simple RF in the visual cortex, which can be expressed by Eq. 11:

$$G(x, y) = \exp\left[-\frac{(x - \mu)^2 + \lambda y^2}{2\sigma^2}\right]\cos[2\pi f(\mu + \phi)]$$

where $\sigma^2$ is the spatial variance, $f$ is the optimal spatial frequency, $\lambda$ is the spatial aspect ratio. The parameter of $\phi \in (-\pi, \pi)$ is a phase offset set that we set $\phi = -\pi/2$ in our investigation. And $\theta \in [0, \pi)$ is the direction parameter of Gabor filter. The response $R(x, y)$ of simple cells to an input image $I(x, y)$ is
calculated by convolution filtering, and the frequency and direction characteristics of IR image can be extracted through convoluting with Gabor function. The expression of convolution filtering is shown in Eq. 12

\[ R(x, y) = G(x, y) * I(x, y) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} G(x-x_r, y-y_r) I(x_r, y_r) \]  

(12)

where \( R(x, y) \) is the output image processed by Gabor function, \( G(x, y) \) is the Gabor function, and \( I(x, y) \) is the input image of dimension \( M \times N \).

LI was first discovered and confirmed by Hartline in 1932 [11]. Dai has proposed a method for simulating the distribution of LI coefficient by exponential function [12], which described the relationship between LI coefficient and the distance between two receptors, and the expression is shown in Eq. 13

\[
\begin{align*}
    h_{mn}(p, q) &= \exp\left(-\frac{d_{ij,pq}}{\rho}\right) \\
    \rho &= \frac{1}{I(x,y)}
\end{align*}
\]  

(13)

where \( h_{mn}(p, q) \) is LI coefficient, \( I(x, y) \) represents the input image, \( d_{ij,pq} \) is the distance between pixel \((i, j)\) and central pixel \((p, q)\) in one inhibition field.

2.2.2. Adaptive gabor filter In this part, we proposed an adaptive Gabor filter. First, the gradient direction of each pixel is calculated according to image information. Then, the direction parameter \( \theta \) is adaptively determined by the calculated gradient directions. Finally, the complete edges corresponding to different directions can be extracted.

Meanwhile, the direction parameter \( \theta \) is determined by Sobel operator in the proposed adaptive Gabor filter. Specifically, the gray value of each pixel in four neighborhoods are weighted, and then the gradient direction is calculated by the difference computation. The partial derivatives \( f_x \) and \( f_y \) of Sobel operator can be determined by

\[
\begin{align*}
    f_x(x, y) &= \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \\
    f_y(x, y) &= \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix},
\end{align*}
\]  

(14)

where \( f(x, y) \) is the gray value of pixel \((x, y)\). Then, the gradient direction angle \( \theta \) of pixel \((x, y)\) can be calculated by Eq. 15.

\[
\theta = \arctan \left[ \frac{f_y(x, y)}{f_x(x, y)} \right]
\]  

(15)

Finally, the Eq. 15 is substituted into Eq. 11, a Gabor filter with adaptive direction parameter is achieved, as shown in Eq. 16.

\[
G(x, y) = \exp\left[ -\frac{\mu^2 + \lambda \nu^2}{2\sigma^2} \right] \cos[2\pi f \mu + \phi]
\]

\[
\begin{align*}
    \mu &= x \cos \theta + y \sin \theta \\
    \nu &= -x \sin \theta + y \cos \theta \\
    \theta &= \arctan \left[ \frac{f_y(x, y)}{f_x(x, y)} \right]
\end{align*}
\]  

(16)
2.2.3. Background prediction based on LI In this part, we used LI network to achieve background prediction. The model of background prediction is shown as follows.

\[
E(x, y) = I(x, y) - \sum_{(p, q) \in S} W(p, q) \Psi(m - p, n - q)
\]  

(17)

where \( E(x, y) \) is the residual value between input image and predicted image, \( I(x, y) \) is input image, \( S \) is the filter window, \( W(p, q) \) is the weight matrix.

Then, the gray value of each pixel in image is regulated according to the residual value \( E(x, y) \) of background prediction, as shown in Eq. 18.

\[
f_{\text{out}}(x, y) = (1 + K \cdot E(x, y)) \cdot f_{\text{in}}(x, y)
\]  

(18)

where \( E(x, y) \) is the residual value, \( f_{\text{in}}(x, y) \) and \( f_{\text{out}}(x, y) \) are the input image and output image. \( K \) is the regulatory factor. Finally, Eq. 18 is substituted into the Eq. 12 to obtain the final output results, and the expression is shown as Eq. 19.

\[
R(x, y) = (1 + K \cdot E(x, y)) \cdot (G(x, y)^* I(x, y))
\]  

(19)

\[
= (1 + K \cdot E(x, y)) \cdot \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} G(x-x', y-y') I(x', y')
\]

where \( R(x, y) \) is the output image, \( E(x, y) \) is the residual value of background prediction, \( G(x, y) \) is the proposed adaptive Gabor filter, and \( I(x, y) \) is the input original image.

2.2.4. Algorithm process Firstly, the gradient direction angle of each pixel is calculated by Sobel operator according to Eq. 15, and the direction parameter \( \theta \) of Gabor filter is determined according to the calculated gradient direction angle. Secondly, LI coefficient is calculated by LI model according to Eq. 13, and the weight matrix \( W \) of background prediction is adaptively determined by the calculated LI coefficient. Then, the predicted image is subtracted from the corresponding input image to obtain the residual value by using Eq. 17. The target or background can be predicted according to the obtained residual value. Finally, the image is processed by adaptive Gabor filter and the residual value of background prediction based on LI simultaneously by using Eq. 19, in which the adaptive Gabor filter is calculated according to Eq. 16.

2.3. KCF tracking algorithm based on multi-channel memory model

2.3.1. Cognitive memory

Memory is a process of encoding, storing and retrieving input information in human’s brain, which can be divided into instantaneous memory, short-term memory and long-term memory [13]. Meanwhile, the human memory system can be described as a Multiple-entry modular memory system, which is composed of multi-dimensional basic cognitive elements [14]. Based on the above understanding, an updating model based on multi-channel memory is established.

As shown in figure 3, the proposed updating model based on multi-channel memory includes a control channel and two executive channels, while each channel contains Instantaneous Memory Space (IMS), Short-Term Memory Space (STMS) and Long-Term Memory Space (LTMS). In the updating model, the IMS stores one template, both the STMS and the LTMS store n templates. The control channel is used for storing the target template. Meanwhile, two executive channels (R-1, R-2) are used for storing the output information of classifier, in which R-1 is used for storing the parameters of the classifier, and R-2 is used for storing the feature of the classifier.
Figure 3. The updating model based on multi-channel memory.

The model mainly updates the template of memory space in the execution channel according to the match between the target template and the template of memory space in the control channel. Meanwhile, the update rules of each channel are different. Specifically, in the memory space of the control channel, the estimation template which obtained from the tracking result of the current frame is firstly memorized in the IMS, and then sequentially match with the templates in the STMS and LTMS. If the match succeeds, the target template in the control channel is updated according to the update rule of the control channel. Meanwhile, the template in the memory space is updated according to the update rules of R-1 and R-2. If the match fails, the estimation templates in the IMS of control channel C and execution channels (R-1, R-2) are memorized in the STMS, respectively.

2.3.2. Algorithm process The process of the KCF Target tracking algorithm based on multi-channel memory model is shown in figure 4.

Figure 4. The process of the proposed algorithm.
The specific steps of the algorithm are as follows:

1. Initialize the multi-channel memory space and target tracking window. Establish the memory space of control channel C and the execute channels (R-1, R-2).

2. In the first frame, the histogram feature \( q_1 \) is calculated by using the initialized target tracking window, and the patch is used as a sample to obtain the HOG feature \( x_1 \) and the parameter \( \alpha_1 \) of the classifier. Meanwhile, these parameters are also regarded as the template and parameters of the next frame’s classifier.

3. In the \( t^{th} \) (\( t \geq 1 \)) frame, firstly, the response of the \( t^{th} \) frame is calculated by using the classifier obtained from the \((t-1)^{th}\) frame according to
   \[
   f(z) = w^T z = \sum_{i=0}^{n-1} \alpha_i \kappa(z, x_i),
   \]
   and calculate the target position. Then, the histogram feature \( q_t \), the HOG feature \( x_t \) and the parameter \( t \) of the classifier template is obtained.

4. The current target histogram feature \( q_t \) is matched with the templates in the STMS of the control channel, and update the multi-channel memory space and the classifier template and the parameter. Finally, obtain the classifier template and the parameters of the \((t+1)^{th}\) frame.

3. Experiments

3.1. IR object detection algorithm based on lateral inhibition and singular value decomposition

The input images are three images with dim targets and area targets. Figure 5 show the detection results. As we can see, the first three methods can enhance the target, yet fail to suppress clutters (figure 5(b)). Top-hat and Shi’s are capable to suppress background, but reserve some clutters, (figure 5(c)). In addition, if the targets tend to be very small, the performance became weak (figure 5(a) and (b)). In comparison, the proposed method not only enhance the targets, but also suppress the clutters well, which makes the proposed method has good detection performance.

![Figure 5. The detection results based on lateral inhibition and SVD.](image)

3.2. IR object detection algorithm based on receptive field and lateral inhibition

The input experiment images and result images of the different methods are also shown in figure 6.
From figure 6, the proposed method has excellent performance on target enhancement and background suppression. According to Column 2, top-hat method works poor in target enhancement when the target is very dim; According to Column 3 and Column 4, the max-mean method and max-median method have poorer abilities on enhancing target and suppressing background compared to the proposed method; According to Column 5, the result of TDLMS has good background suppression ability. However, there are still many clutter backgrounds remained when the backgrounds are complex; According to Column 6, Shi’s method has good performance on extracting small targets from complex background. However, when original image has lots of noises, it remains a lot of noise. According to Column 7, the dim small targets are extracted accurately from complex background and the result images have less clutters and noises, which indicates that the proposed method has excellent performance on enhancing target and suppressing background.

3.3. KCF object tracking algorithm based on multi-channel memory model

Figure 7 shows the experimental results of eight representative sequences from OTB50. It can be seen that compared with the comparison algorithms, the proposed algorithm has higher robustness in the target tracking under complicated conditions for most of the test sequences.
Specifically, the proposed algorithm shows higher robustness in the case of occlusion. For instance, in the results of the Coke sequence, the coke is covered by the plants, and then appears again. According to the results of frame 10, 188 and 236, the comparison algorithms cannot track the coke after it is covered by the plants. While since the proposed algorithm introduces the update model based on multi-channel memory, which can memorize previous scenes. Thus, when the covered target reappears at frame 236, the target matches the template of the memory space in the control channel successfully, which makes the classifier update correctly and continue to accurately track the target.

Meanwhile, high robustness of the proposed algorithm can also be found in the case of target deformation. For instance, according to the results of the Freeman 1, the comparison algorithms IVG, L1APG, MTT deviate from the target after a small gesture transformation of the target (frame 150), and finally loses the target (frame 323). OAB cannot track the target accurately under the conditions of target deformations. The proposed algorithm memorizes target templates in multi-channel memory during the tracking process, so that the classifier can update according to the target’s poses. As a result, it can track the target stably when the target has different gesture changes (frame 40, 150 and 323), which indicates that the proposed algorithm has satisfied adaptability under the conditions of target deformations.

Finally, according to the results of figure 7, the proposed algorithm shows high robustness in the case of background clutters. For instance, the target of video Football is the rugby player’s helmet, which results in the interference of background clutters and similar target. At the frame 220, the helmet of the target is sheltered by that of No. 37 athlete. According to the frame 352, the comparison algorithms IVG, L1APG, MTT tracked the wrong targets (No. 37 athlete) and OAB lost the target when the target athlete collided with the surrounding athletes. By contrast, since the continuous updating of the target template in the control channel of the proposed algorithm, and the classifier of the execution channel is updated accurately, the tracker can accurately track the target under the condition of background clutters and similar target interference.

4. Conclusion
As the most effective bio-intelligence system, Human Visual System (HVS) has significant advantages in image processing. IR target detection and tracking technology has been widely used in transportation, medical, safety and military affairs, etc. However, there still exists some challenges in infrared object detection and tracking, such as dim small target, complex background, heavy occlusion and appearance variation, etc. In this paper, several brain-inspired models and corresponding mathematical models have been studied. Furthermore, the following algorithms are proposed: an infrared target detection algorithm based on lateral inhibition and singular value decomposition, an infrared target detection algorithm based on receptive field and lateral inhibition, an infrared moving dim target detection algorithm based on ALI-PCNN, an infrared target detection algorithm based on GCF-SB visual attention model, a kernel correlation filtering target tracking algorithm based on multichannel memory model, and a robust and efficient discriminative-correlation-filter-based tracking approach based on the Response Map Analysis Network. In a word, applying brain-inspired models and DNNs is beneficial to achieve accurate infrared target detection and robust tracking under complex conditions.

Acknowledgments
The authors would like to thank the Institute of Optics of the University of Rochester (Rochester, NY, USA) and Qingyue Song of the High School Attached to Tsinghua University for their help.

References
[1] Venkateswarlu R 1999 Proceedings of SPIE- The International Society for Optical Engineering 3809 74–83
[2] Yang L, Yang J and Yang K 2004 Electronics Letters 40 1083–5
[3] Yang C, Ma J and Qi S 2015 Applied Optics 54 2255–65
[4] Soni T, Zeidler J R and Ku W H 1993 *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* **2**

[5] Akin O, Erdem E, Erdem A and Mikolajczyk K 2016 *Journal of Visual Communication & Image Representation* **38** 763–74

[6] Shi M, Peng Z and Zhang Q 2011 *High Power Laser and Particle Beams* **23** 906–10

[7] Han J, Ma Y and Huang J 2016 *IEEE Geoscience and Remote Sensing Letters* **13** 452–6

[8] Liu J L and Feng D Z 2014 *Image and Vision Computing* **32** 37–53

[9] Feng X and Milanfar P 2002 *Conference Record of the Asilomar Conference on Signals, Systems and Computers* **1** 478–82

[10] Daugman J G 1985 *Journal of the Optical Society of America A Optics & Image Science* **2** 1160–9

[11] Hartline H K 1938 *American Journal of Physiology* **121** 400–15

[12] Dai S, Liu Q, Li P, Liu J and Xiang H 2015 *Infrared Physics & Technology* **68** 10–4

[13] Wang Y, Eng D O E C, Ying W and Univ, C 2002 *IEEE International Conference on Cognitive Informatics*

[14] Johnson M K and Hirst W 1993 *Mem: Memory subsystems as processes*