Object Tracking Algorithm Based on Adaptive Deep Sparse Neural Network

Lingkang Gu*
School of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui, China
*Corresponding author e-mail:glk_81@163.com

Abstract. Due to the complexity of object tracking easy to produce the tracking drift problem, this paper proposes an object tracking algorithm based on deep sparse neural network. In the particle filter framework, using the Rectifier Linear Unit (ReLU) activation function, according to different situations of object to construct a deep sparse neural network structure, through the finite sample label on-line training, this algorithm can get a robust tracking network. The experimental results show that compared with the current mainstream tracking algorithm, the average tracking success rate and accuracy of algorithm are greatly improved, and according to changes in light, occlusion and fast object movement in complex environment, the algorithm can effectively solve the problem of tracking drift, and show good robustness.

1. Introduction
Object tracking is one of research hotspots of the current and related fields of computer vision, widely used in video surveillance, human-computer interaction and other areas [1-6]. Its purpose is given the initial state of the target in the video (such as location, size, etc.) in the subsequent frame by estimating the target state to realize target tracking.

But in the actual environment, the goal will be to changes in many factors, such as illumination, occlusion, deformation and motion blur [7-12] etc. These algorithms tend to drift or lost target tracking under complex conditions. How to target a more robust, fast tracking, visual tracking is still the current challenging problem.In addition, the traditional tracking method based on single network structure, is easy to lose the original information and is easy to cause the tracking drift to a certain extent[13-15].

To solve the above problems, this paper proposes an online tracking method based on adaptive deep sparse network. We use ReLU activation function [16,17] to structure deep sparse network, which is highly sparse and robust by a limited sample of online training; at the same time as ReLU constant gradient, can effectively solve the "gradient disappear". According to different types of tracking targets, an adaptive multiple network tracking model is constructed, which preserves the structure information of the target better, and has better robustness and adaptability. Experimental results show that the proposed algorithm can not only adapt to complex background and target appearance, scale changes, but also have higher tracking speed.
2. Deep sparse network model

Deep sparse network [16] refers to the deep network structure of only a few nerve cells at the beginning of construction, compared with the traditional neural network, it has better sparseness and robustness, its performance is better in deep structure. Firstly, stacked denoising auto-encoder (SDAE) is trained from a large number of images, and the target information in the image is extracted by using it; then using a sparse representation based dimensionality reduction method, select feature dimension small efficiently from a large number of feature dimension; and then through the introduction of ReLU (rectifier linear unit) activation function, construction depth of sparse neural network model.

2.1. Stacked denoising auto-encoder (SDAE)

SDAE [14] is to learn from millions of images, the reconstruction of the original image, by optimizing the reconstruction error to improve the deep learning network robustness to noise [15]. Using SDAE to extract image features, it will produce a lot of noise or target irrelevant background information. The underlying features of the network are discriminative, which can better describe the changes in the target class, and the high-level features are more semantic. Based on the above consideration, this paper proposes an efficient dimension reduction method based on the sparse representation, to reduce high-dimensional features extracted from SDAE.

Firstly, the first 10 frames of a video sequence are used to construct the dictionary. The construction process of the positive and negative templates in the dictionary is shown in Figure 1. Because the image features which extracted by SDAE are redundant, the sparse representation method is adopted to select the effective features from the mass information in this paper. The formula for feature selection is:

$$\min_s \left\| A^T s - p \right\|_2^{\lambda} + \lambda \left\| s \right\|_1$$

(1)

In the formula, $A \in \mathbb{R}^{K \times (m+n)}$ as dictionary, in which $m$ and $n$ are positive and negative template number, where $m = 10$, $k$ as the feature dimension; $s$ as sparse coefficient vector; $\lambda$ as weighting factor; $p \in \mathbb{R}^{(m+n) \times 1}$ as attribute of each atom, $+1$ as positive template attribute of atom, $-1$ as negative template attribute of atom.

According to the Eq. (1), the sparse coefficient vector $s$ is obtained, and the nonzero elements in $s$ are chosen as the basis of feature selection. The element $S_{ij}$ of $i$-th row and $i$-th column in the projection matrix $S$ is:

$$S_{ij} = \begin{cases} 0 & s_i = 0 \\ 1 & \text{other} \end{cases}$$

(2)

In the equation, $s_i$ is the $i$-th element of the sparse coefficient vectors. By Eq. (2) the dictionary $A$ and candidate sampling $x$ are projected onto a discriminant space, to achieve the goal of feature selection. The dimensionality reduction dictionary $A'$ and candidate state $x'$ can be expressed as:

$$A' = SA, x' = SX$$

(3)

2.2. ReLU activation function

Activation function is the core of neural network, which simulates the activation characteristics of biological neurons, can be used to approximate any nonlinear function. In traditional neural networks, the activation functions used most commonly are the sigmoid function and the tanh function, which have good effects in the feature space mapping of signal.

![Original Image, Negative Templates, Positive Templates](image1)

**Figure 1.** Construction process of the positive and negative templates.
Glorot[16] modified traditional activation functions, and proposed an activation function--ReLU function, which is more similar to the brain neuron model, its function expression is:

\[ \text{ReLU}(z) = \max(0, z) \]  

(4)

2.3. Deep sparse neural network model

Using the ReLU function to build a deep network, after random initialization, about 50% of the hidden layer nodes are suppressed. In order to further improve the sparsity of the network, we add sparse penalty entries in Eq. (2) [17]:

\[ L = \sum_{j=1}^{m_k} \rho(\alpha_j) \]  

(5)

\[ \rho(\alpha_j) = \log(1 + \alpha_j^2) \]  

(6)

where \( m_k \) represents the number of nodes in the \( k \)-th hidden layer, and \( \alpha_j \) represents the activation value of the \( j \)-th node of the \( k \)-th hidden layer, and \( \rho(\alpha_j) \) represents the sparse penalty of \( \alpha_j \). At this point, by solving the optimization problem of Eq. (7), a more sparse deep layer network can be trained. As shown in Eq. (7).

\[ \min_{\theta, x, b, \lambda} \left( \sum_{j=1}^{n} x_j - x \right)^2 + \lambda \left( \sum_{j=1}^{n} \rho(\alpha_j) + \mu \lambda \right) \]  

(7)

where \( \mu \) is as sparse penalty factor.

3. Online tracking based on adaptive deep sparse network

First of all, we divide the tracking target into 3 categories. The classification is based on the initial aspect ratio, \( r = \frac{w}{h} \), where \( w \) and \( h \) represent the width and height of the target respectively. When the \( r \in [2/3, 3/2] \) as the I class object, when \( r \in [0.1, 2/3] \) for II class object, when \( r \in [2, 3/2] \) for III class object. In view of the 3 kinds of tracking targets, 3 different tracking networks are built on the basis of the traditional single tracking network.

3.1. Particle filter algorithm

The particle filter algorithm [18] is a commonly used algorithm in visual tracking. Suppose \( s_t \) and \( z_t \) represent the target states and observations value at the \( t \) moment respectively, the target tracking process can be regarded as the process of finding the maximum possible state at the \( t \) moment according to the observation sequence \( \{ z_{1:t} \} \):

\[ s_t = \arg \max \ p(s_t | z_{1:t}) \]  

(8)

where \( p(s_t | z_{1:t}) \) represents the posterior probability distribution of the target at the \( t \) moment. According to the Bias criterion, we can get:

\[ p(s_t | z_{1:t}) = \frac{p(z_t | s_t) p(s_t | z_{1:t-1})}{p(z_t | z_{1:t-1})} \]  

(9)

The particle filter approaches the posterior probability distribution of the target state by a set of random particle sets \( i=s, w \) with weights. Where the initial weight of particles is \( 1/N \), because particle filter is prone to weight degradation problem, so in the tracking process, through Eq. (10) update the particle weight, and make its sum always maintain 1.

\[ \omega_t^i = \omega_{t-1}^i \cdot p(z_t | s_t) \]  

(10)

\[ \sum_{i=1}^{N} \omega_t^i = 1 \]  

(11)

where the particle filter algorithm is used for random sampling, and \( N \) sampling samples are sent into the tracking network to obtain the confidence of each sampling particle \( z_i = p(s_t | z_{1:t}) \), that is,
posterior probability. Using Eq. (8) we select the largest particle of $\zeta_i$ as the estimated position of the current frame target, that is to obtain the tracking results of the target.

3.2. Network update strategy
The tracking process is subject to the changes of illumination, object occlusion, target deformation and so on. It is prone to tracking drift, and the network parameters must be updated at this time. Network update condition is:

$$\max(\zeta_i) < T || f_n \geq \eta$$

(12)

where $T$ as update threshold, $f_n$ is the last updated cumulative tracking frames, $\eta$ is the largest total number of frames.

3.3. Main steps of this algorithm
Based on the above construction of an adaptive deep sparse network, the main steps of the tracking algorithm are as follows:

Step 1: input image sequence $I_1, I_2, \ldots, I_n$, the initial state of the target $s_0$;
Step 2: select the trace network type (I, II or III) based on $R=w/h$;
Step 3: collect positive and negative samples, and train the tracking network online;
Step 4: for $i=1, 2, \ldots, n$, do:
   (a) $N$ sampling samples are obtained by particle sampling near $(x_{i-1}, y_{i-1})$;
   (b) $N$ sampling samples are fed into the corresponding tracking network to obtain the confidence of each particle $\zeta_i$;
   (c) Using the Eq. (8), the position of the largest particle of confidence is selected as the tracking result;
   (d) Update the network according to Eq. (11);
Step 5: output the target tracking results in each frame, that is, the estimated state $s_0$ of the target;
Step 6: end.

4. Simulation experiment and analysis
To fully verify the effectiveness of our algorithm, we test this algorithm on a PC with Intel i7 3.6 GHz and 32G RAM. The comparison algorithms include: MIL[4], TLD[5], CT[6], DLT[12], OAB[19], and these algorithms use default parameters.

4.1. Quantitative analysis
In this paper, the tracking success ratio criterion is used for quantitative comparison. The tracking success rate is defined as the ratio of the correct tracking frame to all frames of the video sequence. Suppose $E$ is a rectangular box for tracking results, $F$ is the rectangular frame of the target's true position, area( ) is the area of the target tracking state, for each frame, if

$$\frac{\text{area}(E) \cap \text{area}(F)}{\text{area}(E) \cup \text{area}(F)} > 0,$$

then the frame target is correctly tracked. The calculation results are shown in Figure 2. which shows that the performance of this algorithm is better than that of other mainstream tracking algorithms.

![Figure 2. Tracking success rate.](attachment:image.png)
4.2. Qualitative analysis
In the experimental results, we selected 10 groups of representative videos for qualitative analysis. The following qualitative analysis of the algorithm, the analysis results shown in Table 1.

(1) Target deformation: the deformation of non-rigid object tracking. The target is always in motion, and the body of non-rigid deformation in different degree, only this algorithm, MIL and DLT can accurately track the target, and the other algorithms fail to track effectively.

(2) Motion blur: the target movement is too fast and the target area becomes blurred. Because of the rapid movement of the target, except for the algorithm in this paper, the other algorithms have different degrees of target drift. Only the algorithm, CT, MIL and OAB can correctly track the target tracking results, but OAB and MIL appeared deviation.

Table 1. Qualitative analysis results.

| Cases             | Algorithms |
|-------------------|------------|
|                   | MIL        | TLD        | CT         | Our method | DLT       | OAB       |
| Target deformation| √          | √          | √          | √          |           |           |
| Illumination change|            | √          | √          | √          |           |           |
| Scale change      | √          | √          | √          | √          |           |           |
| Rotation          | √          | √          | √          | √          | √          | √          |
| Complex background|            | √          | √          |            |           |           |
| Motion blur       | √          | √          | √          | √          | √          | √          |
| Occlusion         | √          | √          | √          | √          | √          | √          |

In practical application, this algorithm can according to the different types of targets, adaptive selection of different tracking network, realize the tracking of multiple targets, has good robustness in complex environment, effectively solve the problem of target missing and tracking drift. The effect of this system operation is shown in Figure 3.

Figure 3. The effect of this system operation

5. Conclusion
In this paper, an on-line tracking algorithm based on self-adaptive deep sparse network is proposed. The algorithm uses sparse strategy to reduce the dimensionality of high-dimensional features, and improves the computational efficiency. At the same time, in the framework of particle filter algorithm, target localization, occlusion, illumination change, scale change and target fast moving target tracking are completed. Aiming at the problem that a single network cannot take full advantage of the original structure information of an object, an adaptive multiple network selection online tracking algorithm is proposed. Experimental results show that the proposed algorithm has good robustness in complex environment, and effectively solves the problem of target loss and tracking drift.
In the experiment, when the target appears large angle rotation and severe motion blur, the tracking effect is not very good. How to improve the robustness of the algorithm under the target rotation and motion blur is the next focus of the research.

Acknowledgements
This work was supported by “National Natural Science Foundation of China (No. 61300170)”.

References
[1] Zeng Xianyou, Xu Long, Ma Lin, et al. Visual tracking using global sparse coding and local convolutional features. Digital Signal Processing. 2018; 72(1):115-125.
[2] Xu Xiangyang, Li Yuncheng, Wu Gangshan. Multi-modal deep feature learning for RGB-D object detection. Pattern Recognition. 2017; 72(12):300-313.
[3] Ross D A, Lim J, Lin R S. Incremental learning for robust visual tracking. International Journal of Computer Vision. 2008; 77(1-3): 125-141.
[4] Babenko B, Yang M, Belongie S. Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011; 33(8): 1619-1632.
[5] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012; 34(7): 1409-1422.
[6] Zhang Kaihua, Zhang Lei, Yang Minghsuan. Real-time compressive tracking. European Conference on Computer Vision. Florence, Italy, 2012.p.864-877.
[7] Wu Yi, Lim Jongwoo, Yang Minghsuan. Online object tracking: A benchmark. IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013. 9(4): 2411-2418.
[8] Ma Chao, Huang Jiabin, Yang Xiaokang. Hierarchical convolutional features for visual tracking. IEEE International Conference on Computer Vision, Santiago, Chile, 2015.p.3074-3082.
[9] Najafabadi M M, Villanustre F, Khoshgoftaar T M. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2015; 2(1): 1-21.
[10] Li Huanyu, Bi Duyan, Yang Yuan. Research on visual tracking algorithm based on deep feature expression and learning. Journal of Electronics & Information Technology. 2015; 37(9): 2033-2039.
[11] Jin J, Dundar A, Bates J, et al. Tracking with deep neural networks. Annual Conference on Information Sciences and Systems, Baltimore, MD, USA;2013.p. 213-217.
[12] Wang Naiyan, Yeung Dityan. Learning a deep compact image representation for visual tracking. Advances in Neural Information Processing Systems, South Lake Tahoe, Nevada, USA;2013. p. 809-817.
[13] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision. 2015; 115(3): 211-252.
[14] Hou Zhiqiang, Dai Bo, Hu Dan, et al. Robust visual tracking via perceptive deep neural network. Journal of Electronics & Information Technology. 2016; 38(7): 1616-1623.
[15] Glorot X, Bengio Y. Understanding the difficulty of training deep feed forward neural networks. International Conference on Artificial Intelligence and Statistics, Sardinia, Italy; 2010. p. 249-256.
[16] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA; 2011. p. 315-323.
[17] Toth L. Phone recognition with deep sparse rectifier neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada; 2013. p. 6985-6989.
[18] Wang F S. Particle filters for visual tracking. International Conference on Advanced Research on Computer Science and Information Engineering, Zhengzhou, China; 2011. p. 107-112.
[19] Grabner H, Grabner M, Bischof H. Real-time tracking via on-line boosting. British Machine Vision Conference, Edinburgh, Scotland; 2006. p. 47-56.