Chapter 4
Target Tracking Algorithm Based on Visual Perception Mechanism

Peng Lu, Shilei Huang, Chi Liu, Daoren Yuan and Yafei Lou

Abstract A method based on visual perception mechanism is proposed for solving the problem of target tracking. The tracking of target can be achieved in stability. In this paper, the algorithm use neural responses as the visual features. Firstly, the receptive field of cells in primary visual cortex is obtained from natural images. Then the neurons response of background image and video image sequences can be received and calculated the difference, and the difference is compared with dynamic threshold, the target can be detected in this way. Finally, the target tracking can be realized by iterative. Many categories experiment results show that this method improve accuracy and robustness of the tracking algorithm in condition of time-real.

Keywords Target tracking · Visual perception · Overcomplete set · Neural responses

4.1 Introduction

There are a lot of target tracking methods, which are divided into region-based, feature-based, deformable template-based and model-based generally [1], Among them, the typical algorithm include Camshift [2, 3] and SIFT [4, 5], and so on.

The sparse coding model of complete basis requires the orthogonal basis functions [6]. It does not reflect the internal structure and characteristics of images, and also have less sparsity [7]. Overcomplete model more in line with the mechanism of visual feature extraction, and has a good sparse approximation performance [8, 9]. However, the asymmetry of the input space and encode space

P. Lu (✉) · S. Huang · C. Liu · D. Yuan · Y. Lou
School of Electrical Engineering, Zhengzhou University, No 100 Science Road,
Zhengzhou, China
e-mail: lupeng@zzu.edu.cn

Z. Sun and Z. Deng (eds.), *Proceedings of 2013 Chinese Intelligent Automation Conference*, Lecture Notes in Electrical Engineering 256,
DOI: 10.1007/978-3-642-38466-0_4, © Springer-Verlag Berlin Heidelberg 2013

increases the difficulty of the sparse decomposition and the model solution [10, 11].

For the above problems, we use the energy-based models method for solving the overcomplete model, and use the response coefficient matrix instead of the base function matrix for expressing visual features to solve difficulties of the sparse decomposition and the model solution.

4.2 Overcomplete Sparse Coding Model

The sparse coding model is:

$$I = \sum_{i=1}^{m} A_i S_i + N \tag{4.1}$$

where I is a n dimensional natural image, A_i is a basis function with n dimensional vector, N is a Gaussian noise, s_i is the response coefficient, m is the number of basis functions. If $m = n$, formula (4.1) is a sparse coding model of complete basis, if $m > n$, s is a redundant matrix, then formula (4.1) is transformed into overcomplete spare coding model.

We assume that W is receptive field, $A = W^{-1}$ in condition of the model of complete basis. However, A is a redundant matrix in case of the model of overcomplete basis, so it is very difficult to solve A.

To solve the above problems, we use the logarithm of probability density function to define the energy-based models, as following formula (4.2):

$$\log p(x) = \sum_{k=1}^{m} \alpha_k G\left(w_k^T x\right) + Z(w_1, \ldots, w_n, \alpha_1, \ldots \alpha_n) \tag{4.2}$$

where x is a single sample data, n is the dimension of sample data, m is the number of receptive fields, the vector $w_k = (w_{k1}, \ldots, w_{kn})$ is constrained to the unit norm, Z is the normalization constant of w_i and α_i, G is the metric function of the sparsity of neurons response s, and α_i are estimated following with w_i.

In overcomplete basis case, solving the normalization constant Z is very difficult. Therefore, we adopt the score matching to estimate the receptive field. Let us introduce score function which is defined by the gradient of logarithm of probability density function:

$$\psi(x; W; \alpha_1, \ldots, \alpha_m) = \nabla_x \log p(x; w) = \sum_{k=1}^{m} \alpha_k w_k g\left(w_k^T x\right) \tag{4.3}$$

where g is the first-order partial derivative of G.

We used the distance square of score function between parameter model and sample data to get the objective function:

$$\begin{aligned}\tilde{J} =& \sum_{k=1}^{m} \alpha_k \frac{1}{T} \sum_{t=1}^{T} g'\left(w_k^T x(t)\right) \\ &+ \frac{1}{2} \sum_{j,k=1}^{m} \alpha_j \alpha_k w_j^T w_k \frac{1}{T} \sum_{t=1}^{T} g\left(w_k^T x(t)\right) g\left(w_j^T x(t)\right)\end{aligned} \tag{4.4}$$

where $x(1), x(2), \ldots, x(T)$ are T samples.

By the above analysis, the solution process of the receptive field can be summarized as follows: looking for W to promote the objective function to minimize.

We used the gradient descent algorithm to make the objective function minimization:

$$W(t+1) = W(t) - \eta(t) \frac{\partial \tilde{J}}{\partial W} \tag{4.5}$$

where $\eta(t)$ is the learning rate, which changes with time or iteration times.

The algorithm 1 is the learning process of overcomplete set W.

Algorithm 1: Learning of overcomplete set algorithm

Input: Sample images

Output: Overcomplete set W

Steps:

1. Random sampling to the sample images for obtains the training samples;
2. Whiten the samples by the principal component analysis (PCA) method, and project them into whitenization space;
3. Selected the initial vector W_s, and initialize it to the unit vector, set the error threshold ε;
4. Update W according to the formula (4.5), and normalize the unit vector, meanwhile update parameter α;
5. If $norm(\Delta W) \leq \varepsilon$, stop iteration, otherwise, return to step 4;
6. Stop learning, project the learning result W_s back into the original image space, then get the overcomplete set W.

4.3 Target Tracking Algorithm Based on the Visual Perception

Based on visual sparse and competitive response characteristics, only a small amount of neurons is activated to portray the internal structure of images and priori properties [12, 13]. We selected N neurons which have larger response as the visual feature representation of images as shown in Fig. 4.1.

We assume the difference of neurons responds between video sequence image and background image is as follows:

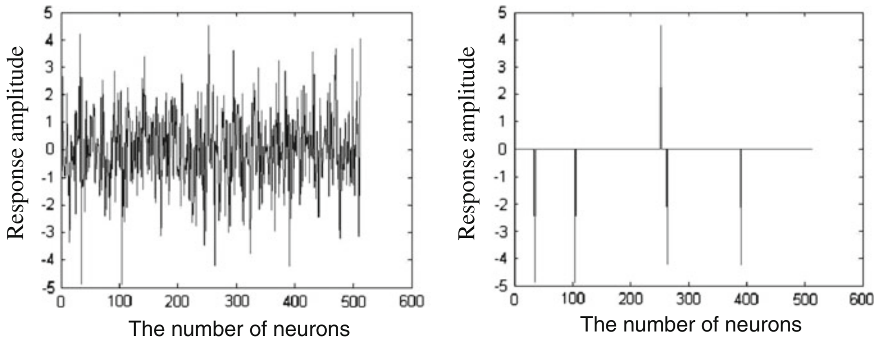


Fig. 4.1 Feature extraction of image visual. **a** The response of neurons caused by image. **b** The representation of visual feature

$$h = \left| s_{vi} - s_{gi} \right| \tag{4.6}$$

where s_{vi} is the response of ith video sequence image patch, where s_{gi} is the response of ith background image patch.

The dynamic threshold is as follows:

$$\delta = \frac{1}{n} \sum_{i=1}^{n} \left| s_{vi} - s_{gi} \right| \tag{4.7}$$

The target tracking algorithm (TTA) is as follows:
Algorithm 2: Target tracking algorithm
Input: Video sequence image and background image
Output: The results of moving target tracking
Steps:

1. Sequential sampling to the video sequence image and background image;
2. Whiten the samples by the principal component analysis (PCA) method;
3. Calculate the neural responses of the video sequence image and background image with the formula $s = Wx$, and take the same number of N largest nerve responses;
4. Calculate the difference h of the neural responses of video sequence image patches and background image patches in the same location, and compared it with the dynamic threshold δ, if $h > \delta$, output the results of the perception, otherwise, no further treatment;
5. Display the recognition results of the target;
6. Then enter the following frame of video sequence, return to step 1.

Flow chart of TTA is shown in the Fig. 4.2.

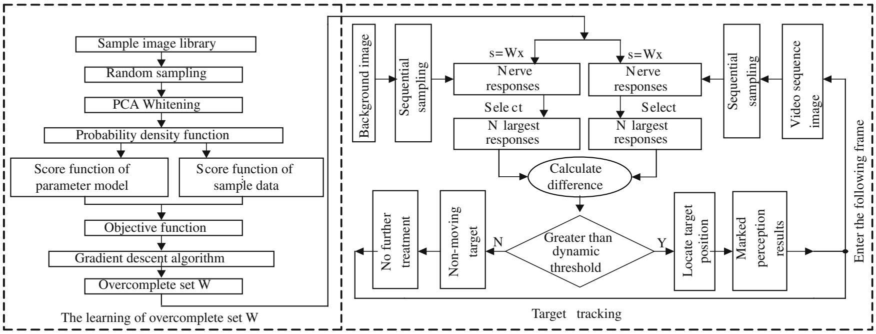


Fig. 4.2 The flow chart of TTA

4.4 Experiment

4.4.1 Learning of Overcomplete Set

Experimental environment: software system-matlab7.0, operating system-Windows XP, CPU-1.86 GHz, memory-1 GB, image resolution-512*512.

Experimental process: Firstly, we select 10 video sequence images and use the 16*16 sliding space sub windows for sampling each image randomly, then we get 5000 16*16 pixels sampling patchess from one image, and 256*50000 sampling data sets from 10 images, and then preprocess the sampling data sets, which is using the PCA method to centralize and whiten the images, and reduce the dimension to 128. The data sets of 128*50000 is dedicated to the input of overcomplete set training. Finally, a overcomplete set representation with 512 receptive fields is estimated based on the energy-based models and the result is shown in Fig. 4.3.

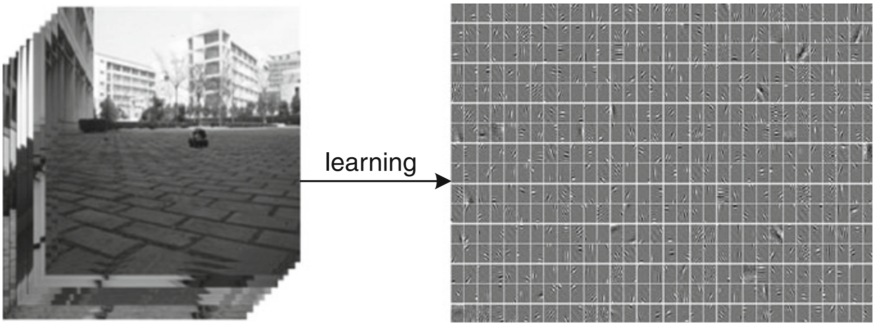


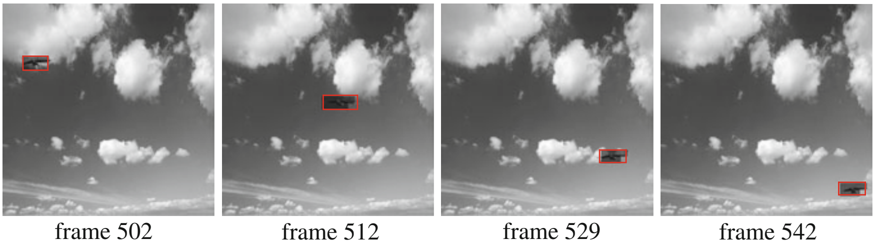
Fig. 4.3 The learning of overcomplete set

4.4.2 Target Tracking

From left to right and top to bottom, we use the 16*16 sliding space sub windows for sampling each image, and get 1024 pixels sampling patches from one image.

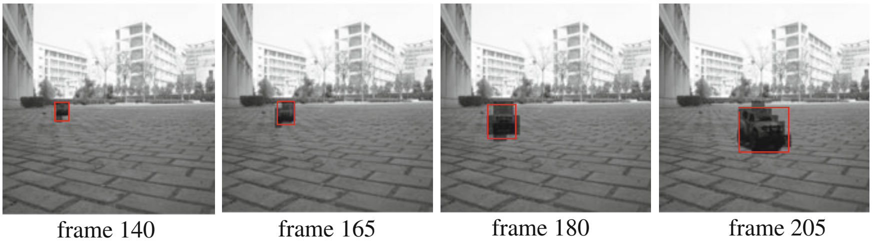
We designed experiments for simple background, target scale change, partial block and complete block. Results of tracking are shown in Figs. 4.4, 4.5, 4.6, and 4.7.

Figure 4.5, the scale and shape of target were changing in the vision. Figures 4.6 and 4.7, the target just passed behind different and similar objects in condition of the partial and complete block, so inter-class change occurs in tracking process.



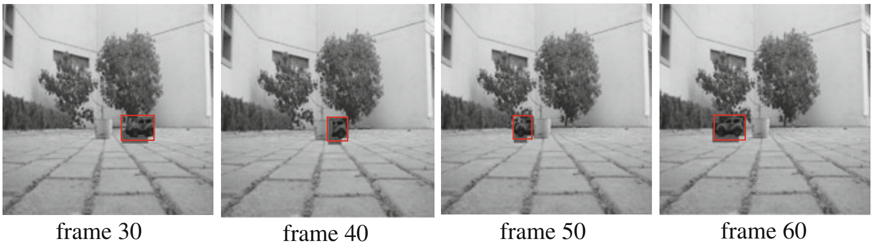
frame 502 frame 512 frame 529 frame 542

Fig. 4.4 Tracking result of the simple background



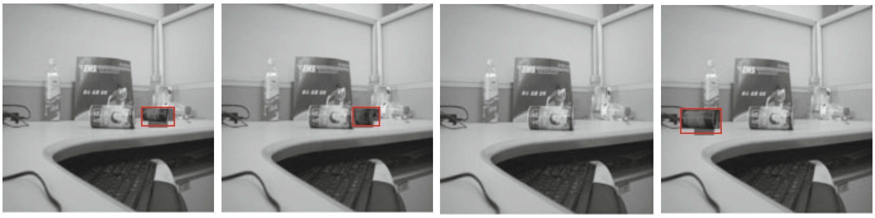
frame 140 frame 165 frame 180 frame 205

Fig. 4.5 Tracking result of the target scale change



frame 30 frame 40 frame 50 frame 60

Fig. 4.6 Tracking result of the partial block



frame 40 frame 45 frame 48 frame 55

Fig. 4.7 Tracking result of the complete block

Table 4.1 The statistics results of three algorithms

	Video sequences (frame)	Right tracking (frame)	False discover target (frame)	False judge non-target (frame)	Wrong tracking (frame)	Recognition (%)
TTA	898	840	26	32	58	93.5
SIFT	898	663	124	111	235	73.8
Camshift	898	564	159	175	334	62.8

Table 4.2 Time-consume comparison of three algorithms

	Video sequence (frame)	Time-consume of one frame (ms)	Total time (ms)
TTA	898	31.2	28017.6
SIFT	898	53.7	48222.6
Camshift	898	18.3	16343.4

In order to verify the validity of TTA, we compared with the typical SIFT and Camshift on the robustness, accuracy and real-time.

4.4.3 Analysis of Results

As can be seen in Figs. 4.4, 4.5, 4.6, and 4.7, TTA which was based on visual perception mechanism achieved tracking of target stably in condition of the block and target scale change. In the Table 4.1, error tracking frames include the false discovery and false judge non-target: the false alarm and missed alarm, the TTA algorithm improves the accuracy of target tracking compared with SIFT and Camshift. It can be seen from the Table 4.2, the time-consume of TTA algorithm is less than the SIFT, and more than the classic Camshift slightly, but to meet the real-time requirement.

4.5 Conclusion

By simulating visual perception mechanism, we established a new kind of target tracking algorithm TTA, and its accuracy and robustness have been improved. TTA algorithm achieved tracking of target stably when occurred scale change of target and block interference, and also target deformation and inter-class exchange at the same time. The furthermore work is we will take further research combined with high-level visual semantics, such as attention and learning mechanism.

Acknowledgments The work for this paper was financially supported by the National Natural Science Foundation of China (NSFC, Grant No: 60841004, 60971110, and 61172152).

References

1. Meng LF, Kerekes J (2012) Object tracking using high resolution satellite imagery. IEEE J Sel Top Appl Earth Obs Remote Sens 5(1):146–152
2. Yin MH, Zhang J, Sun HG, Gu WX (2011) Multi-cue-based CamShift guided particle filter tracking. Expert Syst Appl 38(5):6313–6318
3. Wang ZW, Yang XK, Xu Y, Yu SY (2009) CamShift guided particle filter for visual tracking. Pattern Recogn Lett 30(4):407–413
4. Yao MH, Zhu H, Gu QL, Zhu LC, Qu XY (2011) SIFT-based algorithm for object matching and identification. Remote Sens Environ Transp Eng 271:5317–5320
5. Yu CB, Zhang J, Liu YX, Yu T (2011) Object tracking in the complex environment based on SIFT. IEEE Commun Softw Netw 141:150–153
6. Koldovský Z, Tichavský P (2011) Time-domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space. IEEE Trans Audio Speech Lang Process 19(2):406–416
7. Casaletti M, Maci S, Vecchi G (2011) A complete set of linear-phase basis functions for scatterers with flat faces and for planar apertures. IEEE Trans Antennas Propag 59(2):563–573
8. Mohimani H, Babaie-Zadeh M, Jutten C (2009) A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. IEEE Trans Signal Process 57(1):289–301
9. Labusch K, Barth E, Martinetz T (2009) Sparse coding neural gas: learning of overcomplete data representations. Neurocomputing 72(7–9):1547–1555
10. He ZS, Xie SL, Zhang LQ, Andrzej C (2008) A note on Lewicki-Sejnowski gradient for learning overcomplete representations. Neural Comput 20(3):636–643
11. Hyvarinen A, Hurri J, Hoyer PO (2009) Natural image statistics. Springer, Berlin, pp 289–444
12. Sun H, Sun X, Wang HQ, Li Y, Li XJ (2012) Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. IEEE Geosci Remote Sens Lett 9(1):109–113
13. Dai DX, Yang W (2011) Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geosci Remote Sens Lett 8(1):173–176