International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.8 (2015), pp.333-344
http://dx.doi.org/10.14257/ijsip.2015.8.8.34

Visual Target Tracking Algorithm via Multi-scale Block and Sparse Representation

Ming Li, Cui-Cui Kong, Fu-Zhong Nian and Lei Wang

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, (China)
Email: lim3076@163.com; kongcuicui2010@163.com

Abstract

In this paper, we propose a novel algorithm to deal with the problem of visual tracking in some challenging situations, which is based on sparse representation and multi-scale block. To build target templates, we select distinguishable features between the target and background in each frame of video sequences, dictionary is built by the multi-scale block of target templates. Then, particle filter generates filter distribution in the next frame, the moving target is framed by affine transformation. To describe the current state of the target, we calculate posterior probability for each particle. Finally, the templates are updated online. The experimental results show that the proposed algorithm is superior in accuracy than the classical tracking algorithm, and it has better robustness against to the target posture changes, partial occlusion and illumination variations.

Keywords: *sparse representation, multi-scale block, template update, visual tracking*

1. Introduction

Visual target tracking is one of the important research field of computer vision, which has been widely used in video surveillance, human-computer interaction, video conferencing systems, and the national economy in various fields [1]. As the moving targets are susceptible to suffer from the impact on posture changes, illumination changes and occlusion [2-3], target tracking has some limitations in the application. Among various tracking algorithms, sparse representation is a well-known technique that has attracted a lot of attention.

The first application of sparse representation to visual tracking was proposed by Mei *et al.* [4-5]. Although this method achieves good performance, it takes too much time to solve the l_1 norm minimizations. Another fast object tracking method based on sparse representation is proposed [6], this method was different from the original one in term of image compression block. Block projection for the entire image was sampled from the block diagonal matrix. It lead to slow computing speed and poor quality of image reconstruction. In [7], the confidence map was conducted by the sparse representation coefficient, the conical structure of the confidence map describe the different scales. The method has handled large-scale changes in appearance, but it needs to improve the tracking speed. In [8], each block assigned different weights, sparse solution was calculated by the joint information of target and background. It concentrates on real-time but fuzzy boundaries and the determine of the overlapping blocks. In addition, the global study had also achieved success. Chen F *et al.* [9] have adopted sparse representation appearance model rather than others to enhance the speed of their tracker, but the algorithm have not discussed the tracking frame rates, and the real-time is not high. Venkata I *et al.* [10] have proposed a tracking algorithm based on the similarity projection and Bayesian networks, but the training was complex. Yang yang *et al.* [11] improved the accuracy of single classification by global and local classification, however, this approach failed to track in real-time tracking.

In this paper, a robust visual tracking algorithm based on multi-scale block and sparse representation is proposed. To begin with, we choose a set of target templates, which are used to build multi-scale block of the target dictionary, then paradigm minimize weight is calculated by l_1 regularization in each template dictionary. Finally, the target templates are updated online.

2. Overview of Tracking Models

2.1 Sparse Representation Model

Define a set of template $T=[t_1,t_2,\ldots,\ t_n]\in R^{m\times n}(m>>n)$[12], where t_i is the subset of the training samples from i which is the number of subjects, template set includes n templates. Current tracking results can be represented by linear constraints target template set as follows:

$$y=Ta=a_1t_1+a_2t_2+\ldots+a_nt_n \tag{1}$$

where $a_i=[a_1,a_2,\ldots,a_n]^T\in R^n$ is a sparse coefficient. Considering the change of the target light and noise by the case, tracking results will be estimate as follows:

$$y=Ta+\varepsilon=[T,I,-I][a,e_+,e_-]^T \tag{2}$$

where the vector ε processes noise and occlusion. And positive is represented by e_+, respectively, negative represented by $e\_$ is a small coefficient template, when the target is not in the inter-change.

Research shows that the results of l_1 are more accurate when occlusion occurs. Therefore, the problem can be described as a paradigm track minimization problem:

$$\min\|y-Ta\|_2+\lambda\|a\|_1 \quad s.t. \quad a_i\geq 0,i=1,2,\ldots,n \tag{3}$$

where $a_i\geq 0$ is to ensure the effectiveness of the algorithm. Because of T and y are known, a as the vector coefficients is to be solved corresponding to the dictionary.

2.2 Object Observation Model

Observation model describes the similarity of the target models and the target candidate region. To overcome the difficulties caused by disturbances tracking, target observation model is set up to improve the robustness of the algorithm. For each target candidate region, sparse representation coefficient is calculated to reconstruct target dictionary of the template, reconstructed residual is defined by the candidate area of the observation target, the reconstructed residual likelihood function is defined as follows:

$$p(y_t\mid X_t)=\exp(-\lambda\left\|y_t-\hat{y}_t\right\|) \tag{4}$$

where $\left\|y_t-\hat{y}_t\right\|_2$ represents the reconstruction error between candidate targets and target models. Because $p(\hat{y}_t\mid X_t)$ is equal to w_t^i, the observation likelihood value is equal to the weight of the sample. The smaller the candidate target reconstruction error has, the more reliable the candidate target is.

2.3 Object Motion Model

We denote affine transformation parameters $X_t=(x_t,y_t,s_t,r_t,\theta_t,\lambda_t)$ as target state in frames, where x and y are coordinates of center point, s is change of scale, r is bearing rate, θ and λ are rotation angle and angle of inclination, six affine transformation parameters are mutually independent. Assuming that the state transition model is a Gaussian probability distribution, the state transition probability is obtained by the affine transformation parameters:

$$P(X_{t+1} \mid X_t) = N(X_{t+1} \mid X_t, \sigma) \tag{5}$$

where $N(X_{t+1} \mid X_t)$ is modeled independently by a Gaussian distribution, σ is a covariance diagonal matrix, and the elements of the diagonal matrix are the variance of each affine parameters. $\{X_{t+1}^1, X_{t+1}^2, ..., X_{t+1}^n\}$ is a group of affine parameter sets which are randomly generated in (5). A set of image samples $\{F_{t+1}^1, F_{t+1}^2, \ldots F_{t+1}^n\}$ as a candidate target region are constructed by affine transformation in the current frame. Then, find the area of target from candidate image by sparse representation, which is trained by using previous tracking result.

2.4 Particle Filter

In the particle filter framework, the target motion state is represented by a Gaussian function, which can decide how to select particles. Particle filter algorithm is based on the optimal Bayesian estimation, which is the method of solving the posterior probability. If $p(x_0 / z_0)$ is equal to $p(x_0)$, $p(x_0)$ is the initial probability density function of the known state, period of forecast and update can be expressed as follows:

Forecast: $p(x_k/z_{1:k-1}) = \int p(x_k/x_{k-1}) p(x_{k-1}/z_{1:k-1}) d_{x_{k-1}}$

Update: $p(x_k/z_{1:k}) = \dfrac{p(z_k/x_k)\, p(x_k/z_{1:k-1})}{p(z_k/z_{1:k-1})}$

where $p(z_k / x_k)$ represents the measurement model, $p(x_k / x_{k-1})$ is the state transition model. Posterior probability $p(x_k / z_{1,k-1})$ can be approximated by a recursive weight with a set of samples. After sampling, re-sampling and other operations, we achieve accurate tracking effect.

3. Our Visual Target Tracking Algorithm

3.1 The Structure of the Multi-scale Block Dictionary

Let the original training set consists of n training sample images, which size is $k \times m$. Through the training of multi-scale block, we extend to construct ultra-complete dictionary. Specifically, we construct a grid to divide the original image, the scale of the grid is $l = (0,1,2,\ldots n)$. Each piece of training images is divided into non-overlapping on l different scales. The size of each image block in the original image is on a scale of $1/2^l$, the value of l is 0 to n, the global dictionary is represented by $l = 0$. The number of image blocks on each scale is 2^l, and the total number of image blocks is $2^{l+1} - 2$ on l different scales. All of the sub-picture images in the same position convert into the column vector, and combine together to form a sub-image set. Each sub-block b_i^l on the scale l of rotation for S times in a certain angle, then we obtain a plurality of images $b_{i,1}^l, b_{i,2}^l, \ldots, \ b_{i,s}^l$, where $b_{i,s}^l$ is equal to b_i^l. New training set is trained by these rotated images. Procedure is as follows:

$$B^l = [b_{1,1}^l, b_{1,2}^l, \ldots, \ b_{1,s}^l, b_{2,1}^l, b_{2,2}^l \ldots, \ b_{2,s}^l, \ldots, \ b_{i,1}^l, b_{i,2}^l, \ldots, b_{i,s}^l] \tag{6}$$

$$D = [B^0, B^1, \ldots, \ B^l]$$

where B^l belongs to $R^{(d/i)\times(i\times s)}$. Process is referred to as S times rotation and expansion of the block training set at various scales. D is a training set of $l \times s$ times extended multi-scale blocks through rotation, which is the process of constructing multi-scale block dictionary.

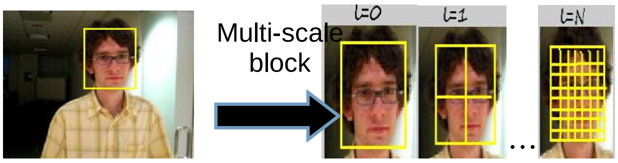


Figure 1. Multi-scale Structure of the Block Dictionary

3.2 The Update of Template and Dictionary

Previous updating algorithms of templates and dictionaries mostly adopt the strategy of overall update [16]. In the tracking process, the target appearance may change due to pose changes, illumination variations and occlusion etc. A fixed appearance model fails to deal with varieties of appearance changes. A new update strategy of the template and dictionary can solve this problem.

For the t-th frame of the video image, assuming we has obtained the target location expressed by X_t through tracking algorithm, the current position of the target need to update the target template. To remove some of the positive samples farther from the target sample, the farther positive samples should be found firstly, then use current location of the target to replace it, we calculate the distance between each positive sample and X_t :

$$d_i = \left\| T^T (x_t - x_i) \right\|_2 \tag{7}$$

where the value of i is 1 to n , T is the initial target template. The furthest distance of sample expressed by k is gained in (7), the sample x_k is replaced by the current sample x_t . To ensure the characteristic continuity between the two frames, a reserve coefficient β is set to update template :

$$x_k \leftarrow \beta x_k + (1-\beta) x_t \tag{8}$$

where the number of β is n . The new positive samples are obtained in (8), and sparse constraints are met in accordance with multi-scale over-complete dictionary block method, template update can be completed.

The dictionary is updated after getting the current state, which is expressed by X_t .The similarity of y and y^* is calculated by the following formula. y^* is a dictionary feature vector of reconstruction, y is the feature vector of X_t . If S meets the threshold condition, y will be updated to the target dictionary:

$$S = 1 - \frac{y - y^*}{\|y\|_2 + \|y^*\|_2} \tag{9}$$

Experimental results show that we obtain better tracking result usually when the value of S is between 0.6 and 0.7. When the threshold condition is met, the following formula calculates the similarity of y and each word in the dictionary, the word of the smallest similarity is replaced by y :

$$\cos\theta_i = \frac{y^T d_p^i}{\|y\| \|d_p^i\|} \tag{10}$$

If the value is too small, you cannot learn target motion state effectively; on the contrary, it will introduce excessive noise which causes the target template dictionary drift, then the tracking is fail eventually.

3.3 The Proposed Algorithm

Firstly, each of the test samples is divided into blocks according to the size of image, and the dictionary represented by D is built by block images of various dimension, which can get the test vector $y = \{y_0, y_1, y_2 \dots y_p\}$. The minimum reconstruction residual error expressed by $\min r^l$ in all categories is solved on the scale of 0. The default threshold is χ, if the value of the minimum reconstruction residual error is greater than χ, then it represents the reconstruction rate is too low, we need to try on a smaller scale; when the value of l is not zero, we can only find the reconstruction residual expressed by $r_{i,k}^l$, where i is the sub-block and k is the class. Reconstruction residual of the entire graph means the average of all the blocks:

$$r_k^l = \frac{1}{2^{2l}} \sum_{i=1}^{2^{2l}} r_{i,k}^l \qquad (11)$$

If the minimum of the reconstruction residuals in all classes is r_k^l, and the value of r_k^l is smaller than χ, and it can be considered that the test image belongs to the class of k; if the value of r_k^l is greater than χ, it can be considered that the scale of this reconstruction error is too large, it needs to try on a smaller scale, the smallest scale reconstruction error is still too large, the system refuses to recognize this image.

Table 1. The Proposed Tracking Algorithm

```
Input: Video frames $\{F_t\}(t = 1,2,\dots,n)$
Output: Target states $\{X_t\}$ $(t = 1,2,\dots,n)$ in frame $F_1, F_2, \cdots, F_t$
        respectively.

for $t = 1 : FrameNumber$ do
    if $t = 1$
    then
        Select the target in the first frame manually. Initialize the appearance
        model using the target observation in the first frame.
    else
        1. In frame $F_t$ ,draw particles $X_t = \sum_{i=1}^{n} w_t^i X_{t-1}$ according to the dynamic
           model $p(x_k / x_{k-1})$.
        2. For each particle $x_k$, calculate the likelihood $p(x_k / z_{1,k-1})$.
        3. Estimate the target state $X_t$ to store the target
           observation $y_t$ simultaneously.
        4. Construct over-complete dictionary by multi-scale block in (6).
        5. Calculation sparse representation coefficient $a_t^i$ with Equation (3).
        6. Determine the status of the target based on the motion model with
           Equation (5).
        7. Update the template and dictionary with the reserve coefficient $\beta$ and the
           threshold $S$ using Equation (8)and (9) individually.
        8. Calculate the reconstructed residual with Equation (11).
    end if
end for
```

4. Experimental Results and Analysis

To evaluate the performance of the proposed algorithm, some standard video sets are tested in Matlab R2012a software environment on an Intel 2.5GHz Dual Core PC with 4GB memory. To further assess the tracking method proposed in this paper, the proposed tracking method is compared with incremental learning algorithm (Incremental Visual Tracking, IVT) [14], the paradigm tracing algorithm (L1 tracker, L1) [5] and multi-instance learning tracking algorithm (Multiple Instance Learning, MIL) [15]. Each method in the four groups is running a video sequence, comparative analysis and tracking results. For each sequence, the initial position of the target is selected manually in the first frame, the initial dictionary is built from the target template through multi-scale block. Tracking results are expressed by rectangular diagram. Observation image uses a window size of 64×64. Particles are used and the template set is updated in every 5 frames to balance the tracking accuracy and the computational efficiency. Both qualitative and quantitative evaluations are demonstrated as follow.

4.1 Qualitative Analysis

Table 2. Our Tracking Image Sequence

Image Sequences	#Frames	Challenging factors
David	462	Illumination changes, posture changes
Deer	71	Abrupt motion, background clutter
Car4	659	Scale changes, illumination changes
Stone	593	Partial occlusion, background complex background clutter

In the David video sequences, the target object undergoes significant illumination change and pose variation. Tracking results are shown in Figure 2. Each image frame is # 21, # 146, # 166, # 261, # 302, # 399. The proposed method can accurately track the target, L1 has no obvious shift. Because the light intensity changes the gray histogram distribution, which can effect histogram features, IVT is sensitive to the impact and gets lost in tracking, the effect of MIL is more obvious.

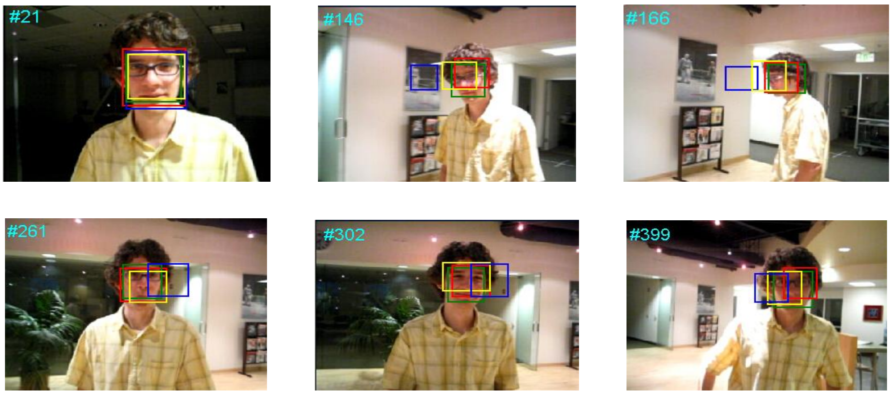


Figure 2. Tracking Results of the David Sequence (IVT is Yellow, MIL is Blue, L1 is Green, our Tracker is Red)

Figure 3 shows the tracking results of the Deer sequences which exhibits challenges on abrupt motion and background clutter. The image frames of tracking results are # 03, # 30, # 35, # 51, # 56, # 71. When deer suddenly leaps up, the proposed tracking algorithm and MIL algorithm show better performance, tracking error appeared at #30 when similar thing occurs, but soon returns to the correct tracking. The methods of L1 and IVT may be more sensitive to the effects.

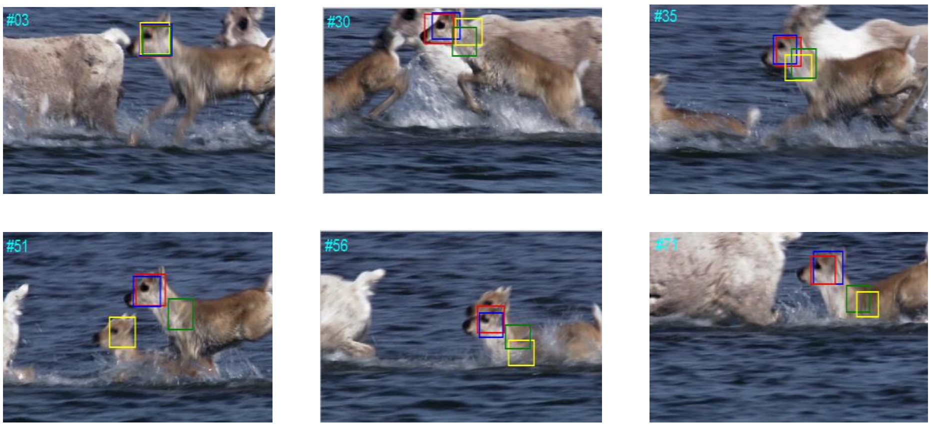


Figure 3. Tracking Results of the Deer Sequence (IVT is Yellow, MIL is Blue, L1 is Green, our Tracker is Red)

In the Car4 video sequences, when the car passes by the bridge and shadow, the light intensity change significantly. Figure 4 shows the tracking result. The image frames of tracking results are # 39, # 197, # 244, # 315, # 399, # 583. When the car passes by the shaded area, larger shift will happen in MIL, goals are not completely lost. L1 appears tracking error at #197, but soon returns to the correct tracking. The method of IVT appears drift more serious after # 197, it will gradually track correctly. Compared with the four methods, the proposed method and L1 perform well.

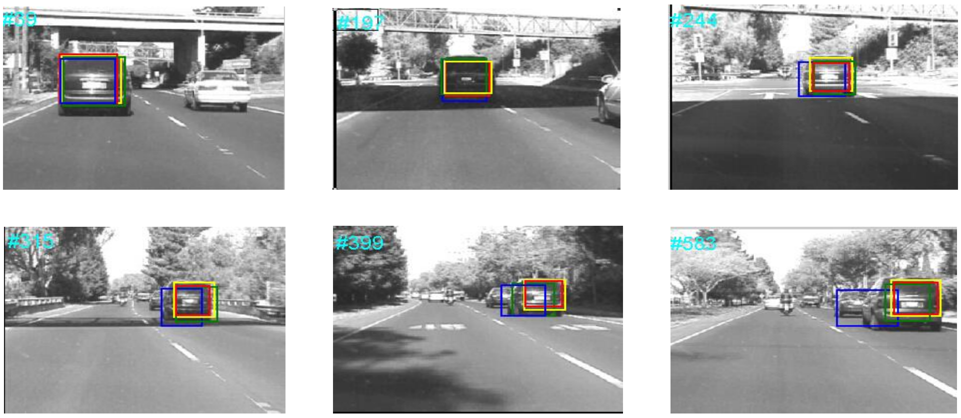


Figure 4. Tracking Results of the Car4 Sequence (IVT is Yellow, MIL is Blue, L1 is Green, our Tracker is Red)

In the Stone video sequences, the target has undergone complex background interference and partial occlusion. The results of tracking are shown in Figure 5. Each frame of the tracking result is # 69, # 135, # 228, # 388, # 405, # 526. When the similar object occurs at #135, the proposed method and L1 can effectively track the target. MIL mistakes analogue for the tracking target, then the tracking targets shift. The method of L1 and IVT appear different shifts, but target is not lost. In comparison, the proposed method performs the best.

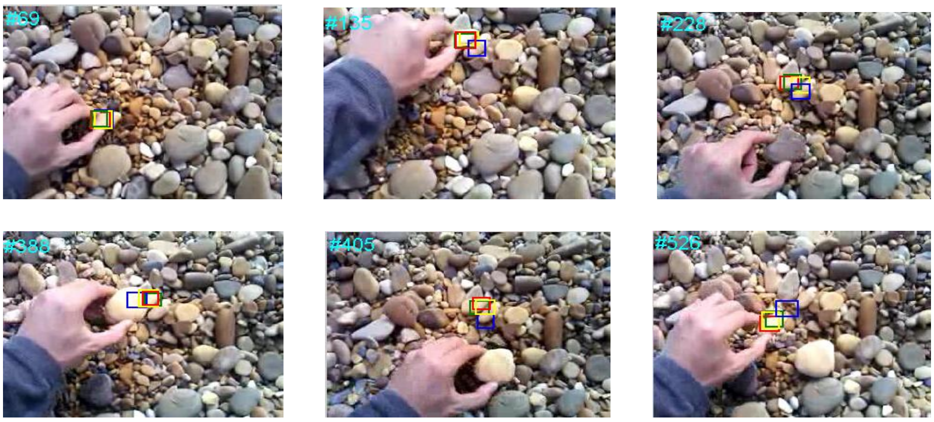
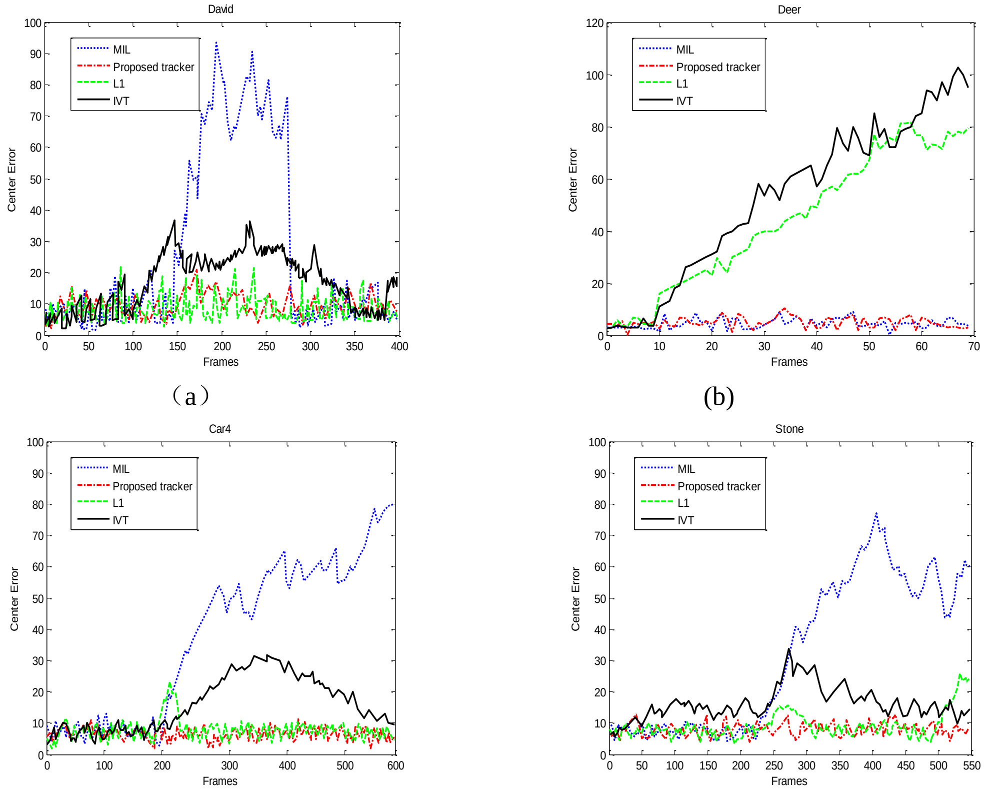


Figure 5. Tracking Results of the Stone Sequence (IVT is Yellow, MIL is Blue, L1 is Green, our Tracker is Red)

4.2 Quantitative Analysis

To further analyze the performance of this algorithm, the algorithm takes center position deviation for quantitative evaluation. Center position error shows the relative positional deviation between the center position and the ground truth, which quantitatively describe the tracking performance between proposed and reference tracker. The results are shown in Table 3 and Figures 6. From the Table 3, Respectively, four tracking methods are used to test the center deviation on 4 video sequences, including the maximum value, mean value and standard value. The values with underline and bold show the best results. The line chart is constituted between center position error and frames in Figures 6.



(a)

(b)

(c)

(d)

Figure 6. Center Error Plot of the Sequences (a) David (b) Deer (c) Car4 (d) Stone

From the Table 3, the proposed method can track the target more accurately in video sequences of David , Stone and Car4 than other methods. We can learn that our tracker and L1 tracker obtain very close and the best results in the Deer sequences; For the David sequences, MIL tracker and our tracker maintain a more stable tracking performance in the light changes; In the Car4 video sequences, our tracking achieve the smallest mean and standard deviation; Our tracker obtain the highest stability in the Stone sequences. Compared with three other trackers, our tracker has the best performance in the David and Car4 video sequences. Taking into account overall performance, our tracker has the best effects in tracking process.

Table 3. Numerical Analysis of the Position Error

	IVT Tracker			MIL Tracker			L1 Tracker			Our Tracker		
	Max	Mean	Std	Max	Mean	Std	Max	Mean	Std	Max	Mean	Std
David	36.27	18.36	10.74	93.28	40.26	29.35	21.22	12.68	5.82	**20.16**	**10.04**	**5.68**
Deer	102.63	58.69	28.52	**8.96**	**5.74**	**3.07**	82.06	39.74	26.81	9.15	5.87	3.29
Car4	31.45	18.49	11.02	80.03	42.82	20.63	22.51	12.01	8.26	**10.26**	**7.65**	**3.96**
Stone	32.44	18.63	8.86	79.98	41.52	25.24	26.71	10.10	**4.98**	**13.22**	**9.83**	5.01

5. Conclusion

This paper presents a robust tracking algorithm via sparse representation and multi-scale block. Based on the multi-scale block processing, the global and local characteristics of the target template image are obtained, paradigm minimize weight is calculated by l_1 regularization in each dictionary. Experiments on four publicly available video sequences indicate that our algorithm performs better robustness than several state-of-the-art algorithms against to illumination variations, posture changes and partial occlusion. However, multi-scale block of the algorithm may lead to a lot of dictionaries, the time of solving convex optimization is too long. The focus of future research work is how to enhance the speed and reduce the computational in this algorithm.

Acknowledgements

This research is supported by the National Natural ScieceFoundation of China (No.612 63019), the Program for International S&T Cooperation Projects of Gansu province (No.144WCGA166) and the Doctoral Foundation of LUT.

References

[1] [1] H. B. Li, D. L. Zeng and Y. Wu, “A two-step multiple targets tracking algorithm based on Mean-Shift and particle filter”, Journal of Chongqing University of Posts and Telecommunications : Natural Science Edition, vol. 22, no. 1, **(2010),** pp. 112-117.

[2] J. Zhang, X. B. Mao and T. J. Chen, “Survey of moving object tracking algorithm”, Application Research of Computers, vol. 26, no. 12, **(2009),** pp. 4407-4410.

[3] Y. F. Liu and J. B. Jiao, “Fast target tracking algorithm based on sparse representation”, Chinese Academy of Science, **(2012)**.

[4] X. Mei and H. Ling, “Robust visual tracking using ℓ 1 minimization”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), **(2009),** pp. 1436-1443.

[5] X. Mei and H. Ling, “Robust visual tracking and vehicle classification via sparse representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, **(2011),** pp. 2259-2272.

[6] J. E. Fowler, S. Mun and E. W. Tramel, "Multiscale block compressed sensing with smoothed projected landweber reconstruction", Proceedings of the European Signal Processing Conference, **(2011),** pp. 564-568.

[7] M. J. Ashwini, R. V. Babu and K. R. Ramakrishnan, "Context-aware real-time tracking in sparse representation framework", Proceedings of the 20th IEEE International Conference on Image Processing(ICIP), **(2013),** pp. 2450-2454.

[8] Y. E. Hou and W. G. Li. "Block sparse representation tracking algorithm with background information", Journal of south China university of technology: Natural Science Edition, vol. 41, no. 8, **(2013),** pp. 21-27.

[9] F. Chen, Q. Wang and S. Wang, "Object tracking via appearance model and sparse representation", Image and Vision Computing, vol. 29, no. 11, **(2011),** pp. 787-796.

[10] I. Venkat, A. T. Khader, K. G. Subramanian and P. D. Wilde, "Recognizing occluded faces by exploiting psychophysically inspired similarity maps", Pattern Recognition Letters, vol. 34, **(2013),** pp. 903-911.

[11] Y. Yang, M. Li and F. Z. Nian, "Vision target tracker based on incremental dictionary learning and global and local classification", Abstract and Applied Analysis, vol. 2013, **(2013),** 10 pages.

[12] J. Wright, Y. Ma, J. Mairal, G, Mairal, G. Sapiro, T. Huang and S. Yan, "Sparse representation for computer vision and pattern recognition", Proceedings of the IEEE, vol. 98, **(2010),** pp. 1031-1044.

[13] Nagarathna, S. Valli and D. Manjunath, "Using an evolution model for efficient estimation and tracking of dynamic boundaries", Proceedings of IEEE Region 10 Annual International Conference, Cebu, Philippines, **(2012)** November 19-22.

[14] D. A. Ross, L. Jongwoo, L. R. Sung and Y. M. Hsuan, "Incremental learning for robust visual tracking", International Journal of Computer Vision, vol.77, **(2008),** pp. 125-141.

[15] B. Babenko, S. Belongie and Y. M. Hsuan, "Visual tracking with online multiple instance learning", IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), **(2009),** pp. 983-990.

[16] D. Wang, H. C. Lu and M. H. Yang, "Online object tracking with sparse prototypes", IEEE Transactions on Image Processing, vol. 22, **(2013),** pp. 314-325.

Authors

Ming Li, he was born in Lanzhou, Gansu Province, China in 1959. He received the bachelor degree in mathematics at Xi'an University of Technology in 1982. Now he is a professor in the School of Computer and Communication at Lanzhou University of Technology, Lanzhou, China. His current research interests are intelligent information processing, pattern recognition, signal processing, face analysis and object tracking. In recent years, he has authored about 30 papers in international journals, national issues and international conference proceedings.

Cuicui Kong, she was born in Liaocheng, Shandong Province, China in 1988. She received the bachelor degree in electronic information engineering at Qingdao Technological University in 2012; she has taken up the subject of Master's Vision Group. Her research interests include intelligent information processing, signal processing and object tracking.

Fuzhong Nian, he was born in Wuwei, Gansu Province, China in 1974. He received the Ph.D. in electronic information engineering at Dalian University of Technology in 2011. Now he is a Associate professor in the School of Computer and Communication at Lanzhou University of Technology, Lanzhou, China. His current research interests are complex system and complex network, intelligent information processing. In recent years, he has authored about 30 papers in international journals, National issues and international conference proceedings.

Lei Wang, she was born in Harbin, Heilongjiang Province, China in 1989. She received the bachelor degree in computer science and technology at Harbin University in 2012; she has taken up the subject of Master's Vision Group. Her research interests include pattern recognition, face analysis and tracking.