Sensors & Transducers, Vol. 177, Issue 8, August 2014, pp. 279-285

Sensors & Transducers

A Novel Object Based Visual Tracking

[1] Gaofeng Li, [1, 2] Wang Lei, [1] Peilong Li
[1] School of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China
[2] Chinesisch-Deutsches Hochschulkolleg, Tongji University, Shanghai, 201804, China
[1] Tel.: +86-182-0195-9628, fax: +86-021-6598-0860
[1] E-mail: 2011gaofengli@tongji.edu.cn

Received: 8 August 2014 /Accepted: 29 August 2014 /Published: 31 August 2014

Abstract: Visual tracking is a challenging task for object with changing appearance in complex scenes. Online learning and detection trackers are developed to resolve the difficulties such as non-rigid, fast motion, occlusion, rotation and scale change. We propose a novel object based tracking method to achieve robust and accurate object. The integrity of object and structural parts are combined for object tracking. Compressive sensing is employed to represent object as root filter. The sparsity of measurement matrix is constrained with superpixel segmentation. The part-based model is adopted to filter the local invariant features of parts. The structural constraint strategy between parts and object is developed for adaptive tracking. We test our proposed algorithm on challenging sequences in real world and make qualitative and quantitative analysis. The experimental results demonstrate the method runs in real time and performs well comparable to state-of-the-art tracking.

Keywords: Compressive sensing, Part-based model, Super pixels, Structural constraints.

1. Introduction

Visual tracking is the problem of generating an inference about the motion of an object given a sequence of video and images [1]. It is widely applied to many fields such as intelligent surveillance, robotics, human computer interaction, action recognition and automatic navigation [2, 3]. State-of-the-art methods are proposed to track object accurately and efficiently, while there are many difficulties to overcome in real-world scenes: illumination variation, occlusion, changes in scale, rotation and deformation. The target is detected and recognized as one object from different angle of view and depth of field with a-priori knowledge. It is deemed as object based perception.

Object tracking method is comprised of three modules: representation, model and update [2], according to which it is categorized as generative or discriminative method. The representation can be points, shape and silhouette of object and the features are extracted with intensity, color and texture. HOG [4] and Haar-like [5] are adopted as template features to compare with candidate targets. The sparse representation is a recent approach to improve tracking performance [6, 7]. Color-based histogram method utilizes colors of mixture distribution to achieve robustness against rotation and partial occlusion [8]. Color is affected by cluttered background, motion blur and human subjective vision. Sevilla-Lara et al. [9] refer to a new representation for object in distribution fields (DFs), which allows smoothing objective function in multiple resolutions. Zhang et al. [10] propose an efficient compressive sensing tracking, which employs the sparse measurement matrix to extract features in compressing space. We combine the sparse compressive representation with object

structure to improve robustness and adaptation of tracking.

The generative method makes a generative and dynamic model to search for similar object. Mean-shift tracker is an iterated optimization approach for searching max probability density distribute of object with kernels [11]. Kalman filter and particle filter trackers are based on the dynamics model of object motion to estimate the posterior probability density of state variables [8, 12, 13]. The bottom-up methods are not adaptive to non-rigid and deformation objects. The discriminative method poses candidates matching as binary classification problem, known as tracking-by-detection. The drift is a main problem in the top-down online learning tracker. MTT induces multi-task learning approach with the interdependencies between particles for avoiding drift [14]. LOT uses locally orderless matching to calculate the likelihood of candidate patches against the noise and changes in appearance [15]. Recently many methods start to combine the two kinds of tracking methods.

In this work we propose a robust object based tracking algorithm to combine generative part model and discriminative detection with compressive sensing. The object is represented in compressed domain and featured by local salient parts and consistent integrality in part-based structure. We improve the sparse representation in the compressive sensing by superpixel segmentation. The part-based model is optimized for adaptive tracking. The object based model is developed to learn and update online.

The paper is organized as follows. In Section 2 compressive sensing is reviewed and improved by superpixels. Section 3 gives a detailed description of object based model. In Section 4 experimental comparisons are performed with our method and state-of-the-art trackers. The conclusions are drawn with contributions and suggestions for future research in Section 5.

2. Compressive Sensing with Superpixels

Compressive sensing is a new method for signal processing, which is efficiently applied to data sampling and reconstruction. When high-dimensional signal x meets the conditions: sparsity, incoherence and RIP (restricted isometry property), x can be projected to low-dimensional space with random measurement matrix [16, 17].

Transform signal x:

$$x = \Psi\alpha\,, \tag{1}$$

where $x \in R^N$, Ψ is the base vector space, α is the sparse representation of x in Ψ domain. The sparsity of x is a-prior condition of compressive sensing. With random linear projecting, the measuring function is given:

$$y = \phi \cdot x = \phi \cdot \Psi\alpha = \Phi\alpha\,, \tag{2}$$

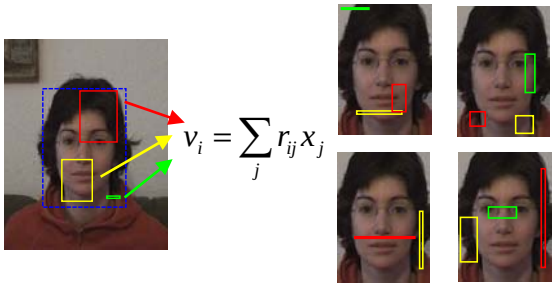
where $\Phi \in R^{M\times N}$ is the random measurement matrix, $M<<N$.

For image random Gaussian matrix satisfies Johnson-Lindenstrauss lemma and RIP [18]. The measurement matrix is selected with [10]:

$$\phi_{i,j} = \sqrt{s} \times \begin{cases} 1, & \text{with probability } 1/2s \\ 0, & \text{with probability } 1-1/s\,, \\ -1, & \text{with probability } 1/2s \end{cases} \tag{3}$$

where $\phi_{i,j} \sim N(0,1)$ for each element of Φ. The sparse representation compresses high resolution image of sequence with little memory.

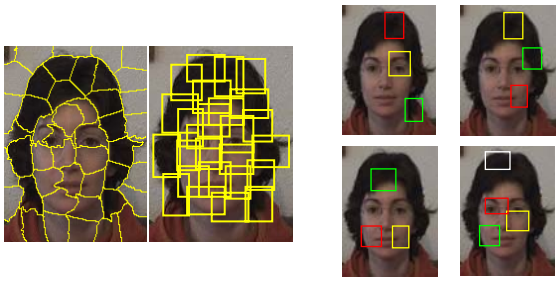
We deploy the configuration of the random rectangles with structural constraints by superpixel method SLIC (Simple Linear Iterative Clustering) [19]. In [10] sparse measurement matrix is made of random Haar-like rectangles for compressing image vector x to low-dimensional vector v (see Fig. 1(a)). In Fig. 1(b) rectangles of group are randomized to satisfy RIP constraint. However, some rectangles have no sense for features extraction in appearance.



(a) Compressive sensing (b) Random Haar-like rectangles

Fig. 1. Compressive representation.

The position and size of random rectangles are constrained according to the features of object by SLIC. The bound box area of object is segmented into superpixel clusters, which are randomly grouped by three or four rectangles to form measurement matrix as shown in Fig. 2. The improved rectangles are closer to salient features of object than above-mentioned compressive representation.



(a) Superpixel SLIC clusters (b) Random SLIC rectangles

Fig. 2. Compressive representation with SLIC.

There are sampling responses of the two compressive representations for one frame in Fig. 3. The red and blue lines respectively denote sampling response curves with the improved rectangles of group and general random rectangles. The red curve is the response of upper right rectangles, which has higher summit and sharper slope than the blue one (upper left rectangles). The proposed sparse measurement matrix preserves the pattern of object with strong discrimination to classify samples.

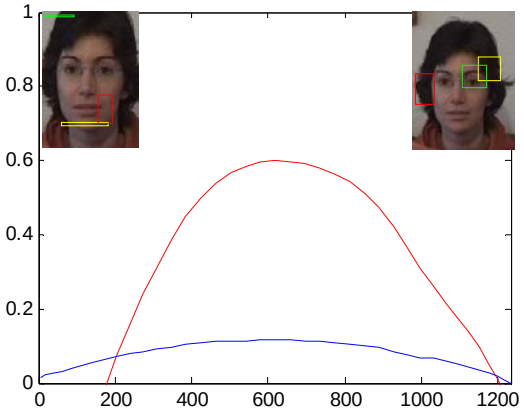


Fig. 3. Response curves of two compressive representations

3. Object Based Model

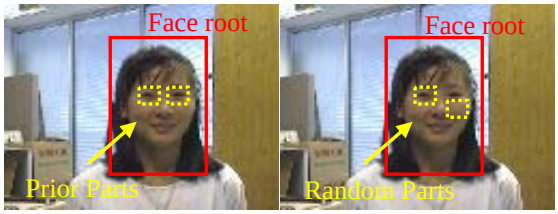
Part model is proposed by Fischler [20] known as pictorial structures, made of local appearance of parts and springs connecting between parts for their location relationship. The object model is represented as a graph $G(V, E)$. The joint $V= (v_1,\ldots, v_n)$ estimates parts location with soft detection of parts features. The edge $E= (v_i, v_j)$ indicates the connection between parts. The optimal target $L^*= (l^*_1,\ldots, l^*_n)$ is given by minimizing energy function:

$$L^* = \arg\min\left(\sum_{1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j)\right), \quad (4)$$

where $m_i(l_i)$ is the matching degree of placing part i at location l_i. $d_{ij}(l_i, l_j)$ is the connecting cost. It is difficult to estimate the location of each part. Thus Felzenszwalb et al. [21] propose the deformable part model (DPM) for global minimizing function and fast detection avoiding unnecessary decisions based on part model. Different from bag of features and HoG template, DPM describes object with each part and parts spatial location constrained to deformable configuration. It has a global root filter F_0 and the spatial part filter F_i, which define the placements restricted to root position with a deformation cost. The placement score is defined the data term plus the spatial term:

$$s = \sum F_i \cdot \phi(H, p_i) + \sum a_i \cdot (\tilde{x}_i, \tilde{y}_i) + b_i \cdot (\tilde{x}_i^2, \tilde{y}_i^2), \quad (5)$$

where $\phi(H, p_i)$ denotes HoG features H in subwindow at p_i. $(\tilde{x}_i, \tilde{y}_i)$ is the location of part i relative to root. a_i, b_i are vector coefficients of a quadratic function. DPM detects object by a collection of parts arranged in a deformable configuration with mixture models to handle significant variation. But it requires offline training which needs much time and a-prior parts as a face made of eyes, nose and mouth. In tracking task object is generally initialized by a bound box. The tracking parts are maybe salient areas not all the time definite components. Therefore, based on DPM we propose object based model for tracking with random SLIC parts not specific known components. As shown in Fig. 4, the parts of DPM are a-prior components with fixed features, while the parts of object based model are not limited to them but selected from salient parts adaptive to tracking.



(a) DPM (b) Object based model

Fig. 4. Root and parts relationship.

The object based model is treated as a simplified combination with root and parts models for tracking. The root model estimates the overall location of the object with compressive representation and Bayes detector. The parts model is made of the rectangles selected from SLIC superpixels. 3 or 5 rectangles are randomly selected from superpixels as parts model. Each part has salient features and a box area with respect to the root similar to DPM [21, 22]. The overall score is the sum of root filtering $H(v)$ and parts energy function $E(p)$. At the initial frame, positive and negative samples are collected for modeling root and parts filters. In sequential frames the root and parts models minimize the score and confirm the position and size of target.

$$Score = H(v) + E(p) \quad (6)$$

The root filter is deployed with Bayesian formula for samples:

$$H(v) = \log\left(\frac{p(y_1 \mid x)}{p(y_0 \mid x)}\right) = \log\left(\frac{\Pi_{i=1}^{n} p(y_1 \mid v_i)}{\Pi_{i=1}^{n} p(y_0 \mid v_i)}\right) \quad (7)$$

where v_i is the set of representation for sample x in compressive domain, y_1, y_0 are the positive and negative samples.

$$H(v)=\log\left(\frac{\Pi_{i=1}^{n}p(v_i\mid y_1)p(y_1)}{\Pi_{i=1}^{n}p(v_i\mid y_0)p(y_0)}\right)=\sum_{i=1}^{n}\log\left(\frac{p(v_i\mid y_1)}{p(v_i\mid y_0)}w_i\right), \quad (8)$$

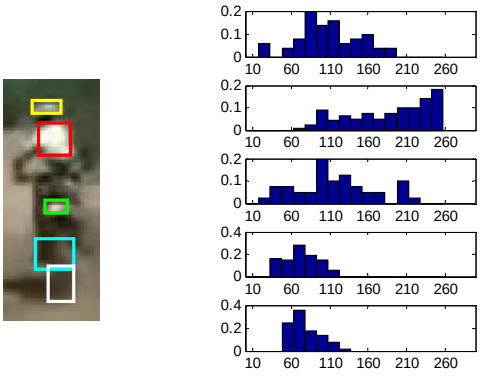
where w_i is the weight of sample at a distance far from origin position. We use it filtering samples to get one coarse object from amount of samples.

The energy function for parts model consists of an appearance term and a smoothing term. The appearance term computes matching degree between part features and its template. The smoothing one penalizes the spring relationship of each other including distance and orientation.

$$E(p)=\sum_{i=1}^{n}F_i\phi(x,y)+\sum_{i,j}^{n}P_{ij}d(x,y), \quad (9)$$

where F_i is each part filter with histogram template, which shows each part matching degree, $\phi(x,y)$ is the feature vector of parts, P_{ij} is the spring relationship between parts, which shows deformable degree, $d(x,y)$ is a set of distance weights of parts. The relationship of parts is considered to restrict P_{ij} according to distance, orientation and shape. In Fig. 5 the parts model of motorcyclist is made of salient parts with histograms of intensity and springs to connect them. The number of parts is flexible and alterable online due to object variation and scenes. When the matching degree of one part is high, it is weighted for model. The two or three parts are enough to track target at a certain time.

According to $H(v)$ the parts model is adjusted to appearance changes. When $H(v)$ is high, the max response of candidate samples is credible. The current root and part features are used to update model. When it is low, the appearance of object might be severely influenced by deformation, occlusion, fast motion or rotation. Parts filter is used to identify parts in large range with the spring of parts. The parts result is fed back to root filter to achieve an optimal target. Learning unit updates parameters of model online adaptive to appearance changes. The global and local features of object, as well as the spatial relationship of them are object oriented tracking conception. It is adaptive and robust for tracking in challenging scenes.



(a) Parts model (b) Histograms of parts

Fig. 5. Motorcycle parts model.

4. Experiments

This section presents experimental results that validate effectiveness and efficiency of the object oriented tracking method (OOT). We evaluate our method on eight challenging sequences used in recent dataset [23] compared with eight state-of-the-art tracking algorithms: Distribution Fields for tracking (DF) [9], Circulant Structure of tracking (CSK) [24], Struck tracking (ST) [25], Compressive tracking (CT) [10], Multi-task tracking (MTT) [14], TLD [26], MIL [6], ASLA [27]. OOT performs robust and adaptive tracking for various situations: fast motion, occlusion, rotation, changes in scale and deformation.

4.1. Qualitative Comparison

OOT is robust to fast motion, which is a common problem leading to motion blur and abrupt shift of camera view. When object abruptly moves partly out of searching range, OOT relocates possible region of object without falling into the local minima. As shown in Fig. 6, there is fast motion and camera shift in Motorcycle sequence. The blue samples are full of search ROI. But the real object is out of the area. The yellow is local minima rectangle. The object is correctly detected as the red one with OOT.

In Woman sequence the partial occlusion appears and the global methods fail to track object. Although only one salient part is credibly detected, the upper parts of woman are tracked by OOT when she is passing by a car shown in Fig. 7. ST also performs well for occlusion in the sequence.

OOT is adaptive to changes in scale with the relative spatial positions of parts. In CarScale sequence, the car varies in scale from far to near and is occluded by trees. See Fig. 8, the light and wheel parts of OOT contribute to track the variations caused by scale and occlusion. TLD and ASLA are adaptive to massive scale changes for multiple scale samples.

Deformable object is difficult to track, especially players, who might suddenly jump, run and rotate out of plane in Basketball sequence (see Fig. 9). OOT avoids drifting with entire features of object and invariant features of parts to track the player.

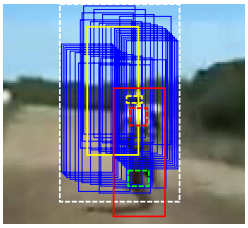


Fig. 6. Fast motion.

Fig. 7. Occlusion.

All results of tracking are shown in Fig. 10, which presents sample frames from sequences with color rectangles denoting algorithms. The sequences

include challenging appearance changes: fast motion, occlusion, rotation, deformation, changes in scale and illumination. ASLA starts to drift at frame 19 in Motorcycle and at frame 191 in Coke. OOT and CSK track well in Bolt sequence. TLD, CT and ST track the moving face accurately in Girl and FaceOcc sequences. OOT, the red one, performs well and scarcely drifts or loses target in the sequences.

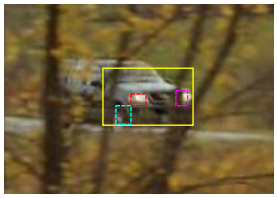


Fig. 8. Scale.

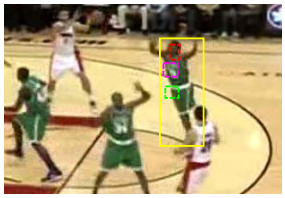
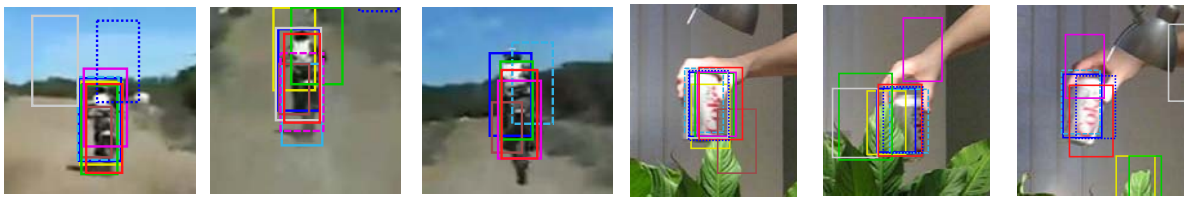


Fig. 9. Deformation.

4.2. Quantitative Results

We calculate the precision and success rate of algorithms for quantitative analysis in Fig. 11. The precision plot is adopted to evaluate the overall performance of algorithm with the center location error given threshold distance of the ground truth. In the precision plots, OOT outperforms other algorithms in CarScale, Woman, Basketball and Bolt. The success plot indicates the ratio of successful frames with the bounding box overlap at different thresholds. OOT performs well for Motorcycle, FaceOcc and CarScale in the success plots. TLD performs accurately in long sequences with P-N learning while it scores low for drifting in deformable and occlusion sequences: Basketball, Bolt and Woman. ST handles well illumination changes and occlusion in Coke, but scale variation. CSK is the top in precision plots of Motorcycle and success plots of Basketball for circulant structure. MIL has poor performance in CarScale, Basketball and Bolt due to the lack of scale and deformable adaptability.

We use the average precision and success of sequences as evaluation metric. Table 1 shows the average of each algorithm for sequences. The average precision of OOT is 0.733. The average success rate of OOT is 0.804. The algorithms are sorted by the total average of precision and success in Fig. 12. OOT scores higher than other algorithms in the sequences evaluation.

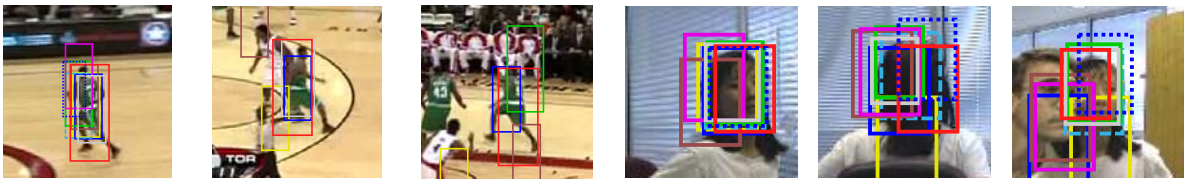


(a) Motorcycle (19, 53, 100)

(b) Coke (52, 191, 291)

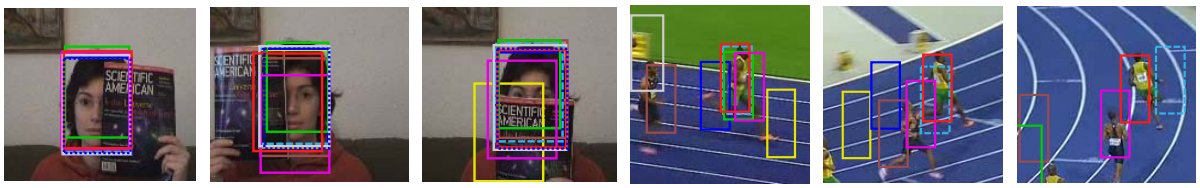
(c) CarScale (141, 168, 243)

(d) Woman (115, 130, 150)



(e) Basketball (126, 487, 725)

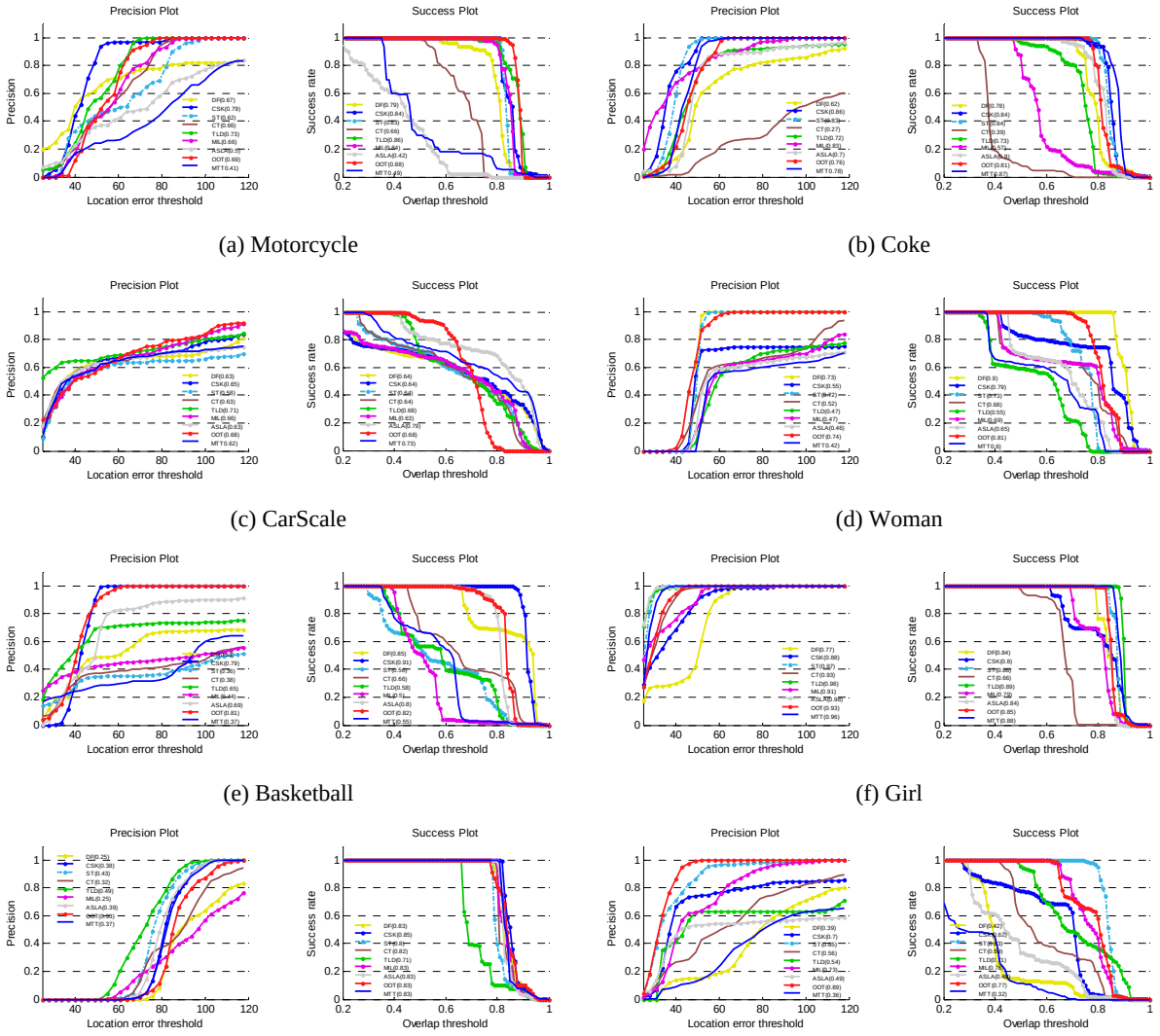
(f) Girl (122, 196, 425)



(g) FaceOcc(188, 388, 864)

(h) Bold (191, 266, 312)

Fig. 10. Sample frames from sequences (Number in brackets is frame number).



(a) Motorcycle (b) Coke

(c) CarScale (d) Woman

(e) Basketball (f) Girl

(g) FaceOcc (h) Bolt

Fig. 11. Precision and success rate of algorithms.

Table 1. Precision and success of algorithms.

	Precision	Success	Average
DF	0.578	0.757	0.668
CSK	0.706	0.788	0.747
ST	0.652	0.757	0.705
CT	0.577	0.648	0.613
TLD	0.686	0.728	0.707
MIL	0.615	0.s698	0.657
ASLA	0.586	0.698	0.642
OOT	**0.733**	**0.804**	0.768
MTT	0.556	0.661	0.608

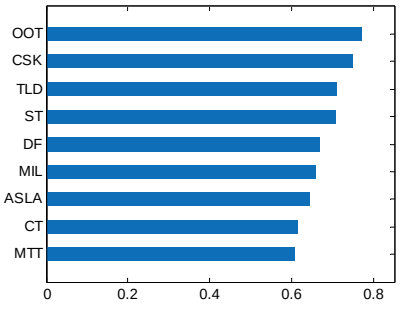


Fig. 12. Total average of algorithms.

5. Conclusions

Object based visual tracking is a robust tracking algorithm and adaptive to appearance variations of object in challenging scenes. OOT represents object with compress sensing improved by superpixels, which have high efficiency with sparsity and local features for enormously sampling. OOT shows advantages in difficult situation and prevents the search from getting stuck in local minima with root and parts models, which solves ambiguity and sensitivity to constrained spatial structure. Experiments demonstrate OOT is an efficient and effective method and it performs well in terms of precision and success rate comparable to state-of-the-art trackers. The future work is intended to improve compressive features space representation and study background model in OOT.

Acknowledgements

The authors acknowledge the support by National High-tech Research & Development Program (863: No. 2012AA7041003).

References

[1]. D. A. Forsyth, J. Ponce, Computer vision: A modern approach, *Prentice Hall, Alan Apt*, NY, 2002.

[2]. A. Yilmaz, O. Javed, M. Shah, Object tracking: A survey, *ACM Computing Surveys*, Vol. 38, Issue 4, 2006, pp. 1-45.

[3]. S. Salti, A. Cavallaro, L. D. Stefano, Adaptive appearance modeling for video tracking: survey and evaluation, *IEEE Transactions on Image Processing*, Vol. 21, Issue 10, 2012, pp. 4334-4348.

[4]. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, USA, 20-26 June 2005, pp. 886-893.

[5]. P. Viola, M. J. Jones, Robust real-time face detection, *International Journal of Computer Vision*, Vol. 57, Issue 2, 2004, pp. 137-154.

[6]. B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, Issue 8, 2011, pp. 1619-1632.

[7]. X. Mei, H. Ling, Robust visual tracking and vehicle classification via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, Issue 11, 2011, pp. 2259-2272.

[8]. K. Nummiaro, E. Koller-Meier, L. Van Gool, An adaptive color-based particle filter, *Image and Vision Computing*, Vol. 21, Issue 1, 2003, pp. 99-110.

[9]. L. Sevilla-Lara, E. Learned-Miller, Distribution fields for tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, USA, 16-21 June 2012, pp. 1910-1917.

[10]. K. Zhang, L. Zhang, M.-H. Yang, Real-time compressive tracking, in *Proceedings of the 12th European Conference on Computer Vision*, Florence, Italy, 7-13 October 2012, pp. 864-877.

[11]. D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, Issue 5, 2003, pp. 564-577.

[12]. Y. Youngrock, A. Kosaka, A. C. Kak, A new Kalman-filter-based framework for fast and accurate visual tracking of rigid objects, *IEEE Transactions on Robotics*, Vol. 24, Issue 5, 2008, pp. 1238-1251.

[13]. X. Bai, Q. Li, T. Zhang, et al, A novel ball tracking method using dynamic Kalman filter with two-stage ball search, *Journal of Computational Information Systems*, Vol. 9, Issue 9, 2013, pp. 3373-3381.

[14]. T. Zhang, B. Ghanem, S. Liu, et al, Robust visual tracking via multi-task sparse learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, USA, 16-21 June 2012, pp. 2042-2049.

[15]. S. Oron, A. Bar-Hillel, D. Levi, et al, Locally orderless tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, USA, 16-21 June 2012, pp. 1940-1947.

[16]. D. L. Donoho, Compressed sensing, *IEEE Transactions on Information Theory*, Vol. 52, Issue 4, 2006, pp. 1289-1306.

[17]. E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Transactions on Information Theory*, Vol. 52, Issue 2, 2006, pp. 489-509.

[18]. R. Baraniuk, M. Davenport, R. Devore, et al, A simple proof of the restricted isometry property for random matrices, *Constructive Approximation*, Vol. 28, Issue 3, 2008, pp. 253-263.

[19]. Appu R. S. Achanta, Kevin Smith, Aurélien Lucchi, Pascal Fua, Sabine Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, Issue 11, 2012, pp. 2274-2282.

[20]. M. A. Fischler, R. A. Elschlager, The representation and matching of pictorial structures, *IEEE Transactions on Computers*, Vol. 22, Issue 1, 1973, pp. 67-92.

[21]. P. Felzenszwalb, D. Mcallester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Alaska, USA, 23-28 June 2008, pp. 1-8.

[22]. P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, et al, Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, Issue 9, 2010, pp. 1627-1645.

[23]. Y. Wu, J. Lim, M. H. Yang, Online object tracking: a benchmark, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland USA, 23-28 June 2013, pp. 2411-2418.

[24]. J. F. Henriques, R. Caseiro, P. Martins, et al, Exploiting the circulant structure of tracking-by-detection with kernels, in *Proceedings of the 12th European Conference on Computer Vision*, Florence, Italy, 7-13 October 2012, pp. 702-715.

[25]. S. Hare, A. Saffari, P. H. S. Torr, Struck: Structured output tracking with kernels, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 263-270.

[26]. Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, Issue 7, 2012, pp. 1409-1422.

[27]. X. Jia, H. Lu, M. H. Yang., Visual tracking via adaptive structural local sparse appearance model, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, USA, 16-21 June 2012, pp. 1822-1829.