ICIC Express Letters
Part B: Applications
Volume **9**, Number **2**, February **2018**

ICIC International ©2018 ISSN 2185-2766
pp. 163–170

MODELING VISUAL APPEARANCE AS A GLOBAL OPTIMIZATION PROBLEM IN VISUAL TRACKING

Longkui Jiang[1], Chuntian Bai[2], Yuru Wang[2] and Hongguang Sun[2]

[1]School of Information Engineering
Jilin Business and Technology College
No. 1666, Kalun Lake Ave., Changchun 130507, P. R. China
jlongkui@163.com

[2]School of Computer Science and Information Technology
Northeast Normal University
No. 2555, Jingyue Street, Changchun 130117, P. R. China
wangyr915@nenu.edu.cn

Received August 2017; accepted November 2017

Abstract. *Visual model is a key factor in visual tracking problem because its discriminative ability determines the robustness and stability of the tracker directly. This paper made major research on developing a dynamic appearance model which is optimal for the target along the whole sequence. Specifically, the problem of constructing a visual model adaptive to variable environment is resolved as a global optimization problem. The discriminative ability of the appearance model is enhanced in two-fold. First, the target is represented in a hierarchical local patches based visual model, and multiple predefined visual cues are integrated to model target's complex appearance. Second, an evaluation scheme is exploited to define the discriminative ability and an optimal model is designed to realize adaptiveness. The proposed model is tested on challenging video sequences from the PAMI 2016 tracking Benchmark, and is validated to be robust and effective.*

Keywords: Image processing, Visual tracking, Global optimization, Visual model

1. Introduction.

Visual tracking has received much attention in recent years due to its potential value theoretically and practically in the fields of intelligence video surveillance, self-driving vehicles, and robotics and so on. Despite that much progress has been made in recent years, developing sufficient robust tracking system is still an open problem in real applications. The main challenges lie in the two aspects, including complex tracking conditions and object variable appearance. In particular, the difficulties include illumination changes, occlusions, structural variation, cluttered scenes, fast motion, etc. The direction of handling the above difficulties focuses on the two aspects of constructing a robust visual model and developing tracking algorithms.

As a basic problem in visual tracking, visual model is a vital issue and determines the robustness and effectiveness directly. In this field, a variety of well-known features have been developed. Overall, the existing visual models can be categorized into two classes, the global model and local model.

The global model refers to the visual features of the object region [1]. The famous model includes color histograms [2] or attributes [3], subspace-based features [4], Haar-like features [5], LBP (Local Binary Patterns) [6], HOG (Histogram of Oriented Gradient) [7], SIFT (Scale-Invariant Feature Transform) [8], SURF [9], covariance matrix [10], 3D-DCT [11], shape features [12], etc. In addition, many researchers combine several complementary cues. In detail, multiple visual cues are integrated in the manner of weighted sum, multiply, hierarchical or min-max, selection, D-S rules. The above global models have been demonstrated to be effective in some specific tracking conditions. However, they are low-level features and not tuned to variable tracking conditions. In many real applications, the target object always shows variable appearance caused by environments influence or self-structural changes. For such cases, the global model shows less robustness.

In comparison with global visual model, the local model shows more robustness to target appearance changes. Many local models have been developed. Kolsch and Turk [13] proposed a flock-of-features model, where the target is represented as local patches. Hoey [14] extended his model. This model shows disadvantages when target shows local abrupt motions. Yin and Collins [15] proposed to constrain model variation by the global affine transformation. Besides, Martinez and Binefa [16] proposed a kernel triangle connection method, and Chang et al. [17] employed MRF (Markov Random Filed) to constrain the structure among patches. Similar models contain star model [18]. Cehovin et al. [19] introduced global feature into local model for the first time. Indeed, both the global and local features are important in modeling target's appearance. The reason lies in the fact that, in real tracking condition, on the one hand, the target appearance will take global variance influenced by environmental changes; on the other hand, it will show local distortion or warping. That is to say, in order to provide a robust representation to target variable appearance, both the global and local features should be extracted.

Since 2013, many researchers have focused on deep learning based visual model. Considering both the foreground and background information, these tracking methods have excellent tracking results in many tracking problems with complex scenes, but the main difficulty is the lack of training data. As we all know, deep model succeeds by learning a large number of labeled training data, but target tracking provides only the first frame of Bounding-box as training data. At present, the deep learning based method solves the problem by on-line fine-tuning of tracking data, and the typical methods include auto-encoder [20], CNN (Convolutional Neural Network) [21-23], RTT (Recurrently Target-Attending Tracking) [24]. The main problem facing the current methods is the suitability of off-line pre-training for on-line tracking and the sensitivity of online updating to fine-tuning. Basically, deep learning based visual model extracts both global and local features and it also belongs to the discriminative model. The discriminative considers both the foreground and background, and will show more robustness than the generative model. Therefore, this paper will focus on the discriminative model.

Another key problem in appearance modeling is the updating. In order to keep the visual model updating with the target variations during tracking, designing an effective online updating scheme is a crucial point. Seldom research is made directly in model updating. Some model employed in discriminative model based tracking method as Adaboost [7], similar updating scheme is employed. Specifically, for each weak classifier, the feature is boosted, and with the tracking going on, the classifiers are updated by new coming samples. For many tracking algorithms like particle filter, this scheme is not applicable.

In this work, inspired by the mechanism of human vision, we made major research on constructing a robust and adaptive local patches based visual model. The main idea behind the proposed model is to employ genetic algorithm to optimize a layered visual model. We develop a hierarchical local patches based visual model, so as to adapt with variable conditions. Moreover, the local patches are integrated in weighted sum way, in order to model their different discriminative abilities. Specifically, the integration parameters are looked as optimization parameters in a global optimization problem, and resolved by GA (Genetic algorithm).

The advantages of the proposed layered visual model are three-fold.

(1) Like human vision, the proposed visual model is constructed in a hierarchical way. Both local and global features are extracted.

(2) The hierarchical cues are weighted integrated, and are tuned with the tracking conditions variation.

(3) The visual model is updated in an optimization model, and the parameters are set to be optimal at each video frame.

The rest of this paper is organized as follows. Section 2 introduces the hierarchical patches model in particle filter framework. Section 3 presents the optimization model and gives the definition of model evaluation. Experimental results are given in Section 4. Finally, we conclude this work in Section 5.

2. **Layered Patches Based Visual Model.** For human vision, when dealing with object tracking, the features in different scales are extracted. And with the tracking going on, their weights take changes. In this paper, the target is represented as patches in different scales, and assigned with updating weights.

In particle filter, target tracking is resolved as a state estimation problem in a Bayesian framework. Given the observations of the object $y_{1:t} = [y_1, y_2, \ldots, y_t]$ up to time t, the destination of particle filter based tracking is to estimate the posterior distribution $p(x_t|y_{1:t})$ of target. In Bayesian point of view, it is resolved as:

$$p(x_t|y_{1:t}) \propto p(y_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \tag{1}$$

where $p(x_t|x_{t-1})$ is the dynamic model, $p(y_t|x_t)$ is the observation model, and $p(x_{t-1}|y_{1:t-1})$ is the prior distribution. Usually, the posterior distribution is difficult to be obtained, and it is resolved by Monte Carlo sampling in particle filter. $p(x_t|y_{1:t})$ is approximated by a set of weighted samples $\{x_t^i, w_t^i\}_{t=1}^N$, and the target state is usually determined by the optimal or the mean of the particles.

In the hierarchical model, the target is represented in a layered patches model. In detail, the target region is represented as a set of patches with different sizes. In the first level, the whole region is divided into patches of small size $n \times n$. In the second level, it is divided into 4×4 components; and in the third level, 2×2 components; finally, the whole region is deemed as one patch. In particular, each patch in the hierarchical model, we extract three visual features including HSV and HOG. Then, the target is observed locally and globally, and represented by features in different scales.

Each patch leads to one particular cue in the integration model. Over all, the observation model is integrated by m cues as $y_t = (y_t^1, y_t^2, \ldots, y_t^m)$, and $p(y_t|x_t)$ is the joint likelihood value of m cues.

$$p(y_t|x_t) = \prod_{i=1}^{m} p\left(y_t^i|x_t\right) \tag{2}$$

where $p(y_t^i|x_t)$ is usually calculated as a similarity measured in distance like

$$p\left(y_t^i|x_t\right) = \kappa_t\left(y_t^i, T_i\right) \propto e^{-d_t^2\left(y_t^i, T_i\right)/\sigma^2} \tag{3}$$

where T_i is the template of feature i, and $d_t^2\left(y_t^i, T_i\right)$ is the distance from the observation to the template. Different cues will show variable confidence or discriminative ability with the tracking going on. Therefore, they should be assigned with online updating weights, so as to capture target's variable appearance. Then, the joint likelihood value becomes:

$$p(y_t|x_t) = e^{-\sum\limits_{i=1}^{m} \pi_t^i d_i^2\left(y_t^i, T_i\right)/\sigma^2} \tag{4}$$

where $\{\pi_t^i\}_{i=1}^m$ is the online updated weights, and they are usually constrained by $\sum_{i=1}^m \pi_i = 1$.

3. **GA Based Online Model Optimization.** The settings of integration parameter will influence the robustness of the visual model. They are updated online for adaptiveness. At each video frame, there must exist a best setting. However, when the integration cues are of large number, the large range of values causes the exact solution time consuming. Therefore, an approximated optimal solution is plausible. Then, how to transform

the weights setting problem to an optimization problem? In this section, we define the integration parameters setting problem with the tracking confidence. Particularly, a random samples based model is employed to define the confidence.

The samples in Monte Carlo sampling are reused in this procedure to evaluate model confidence. On the premise that the feature with more discriminative ability possesses more confidence, its confidence with specific parameters setting is reflected by the observation distribution of the samples. If a model with sufficient discriminative ability is employed, the observed values of these samples will show an approximate unimodal distribution like a Gaussian distribution. Specifically, the samples lying in the target and the backgrounds regions should be easily departed from each other, in another word, of large margin between the two classes. Our goal is not to optimize the classification margin, but to optimize the best positive and negative samples set.

3.1. **Optimization model definition.** For a given integration parameters (here, the cue weights) setting, rank the samples by their weights, and extract two sample sets s_o and s_b, where s_o represents the top n_o samples with higher weights, and s_b represents n_b samples with lower weights. These two sets are representative for the target and background samples.

To find the optimal integration parameters setting is to resolve the following optimization problem:

$$\begin{aligned} &\min f(\pi_t) = \left(\sigma_t^o \sigma_t^b\right) / (m_t d_t) \\ &\text{s.t. } \sum_{i=1}^{m} \pi_i = 1 \end{aligned} \tag{5}$$

where σ_t^o and σ_t^b are defined as $\sigma_t^o = \frac{1}{n_o} \sum_{i=1}^{n_o} \left|x_t^i - \mu_t^o\right|$ and $\sigma_t^b = \frac{1}{n_b} \sum_{i=1}^{n_b} \left|x_t^i - \mu_t^b\right|$, and represent the within class distances. $m_t = \frac{1}{n_b} \sum_{i=1}^{n_b} \left|x_t^i - \mu_t^b\right|$ and $d_t = \left|\frac{1}{n_o} \sum_{i=1}^{n_o} d_t^i - \frac{1}{n_b} \sum_{i=1}^{n_b} d_t^i\right|$ represent the inter-class distances.

3.2. **Biological evolutionary algorithm based integration parameters optimization.** For the above optimization problem, when the problem scale is small, it is easy to calculate an exact solution. For the presented layered visual model, the exact solution is time consuming, because examining all possible solutions of a specific problem is virtually infeasible. Therefore, biological evolutionary algorithm is optional for this problem. Specifically, GA is employed in this paper to approximate the optimal solution. GA is a search procedure within a problem's solution domain, and it offers an optimization heuristic inspired by biological natural selection.

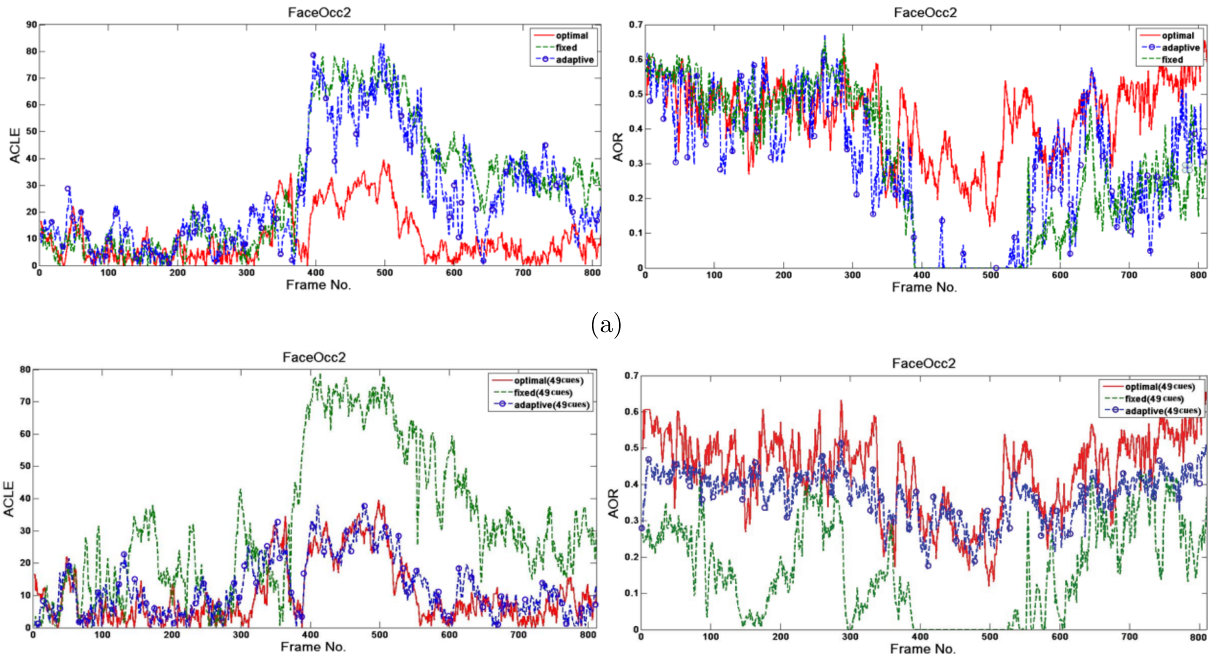
In GA terms, a solution to the problem is represented as an individual "organism" of a large population. Essentially, a GA attempts to reach an optimal solution by mimicking the processes of natural selection and evolution. In each iteration, the entire population is replaced by the many offspring created by the crossover operation. GA starts from a fixed-size population of randomly generated solutions. In each iteration, the entire population is evaluated using the fitness function defined as Formulation (5). The successful performance of a GA depends mainly on the appropriate choice of chromosome representation, crossover operator, and fitness function. The chromosome representation and crossover operator should yield an enhanced solution by merging two "promising" chromosomes that are passed on to the next generation.

4. **Experiments and Analysis.** The proposed method is tested on the PAMI 2016 visual tracker benchmark [25], and ten challenging videos are chosen for further analysis. And the challenges are classified into two groups including occlusion, and complex backgrounds. For illustration, two types of observation models are employed, three-cues and 49-cues (a pyramid model like [26]). Specifically, the observation model including three

cues (HSV, HOG, and LBP) is employed to demonstrate the approximation ability of GA to accurate solution in optimizing weights. And the observation model employing 49 cues is employed to demonstrate the effectiveness of our method in real applications. And in our experiments, the parameters of GA are set as follows: the population number is 100, the cross rate is 0.8, the mutation rate is 0.4, and the mutation factor is set to be 0.01.

4.1. **Occlusion.** Occlusion is a challenging problem for visual tracking because the severe appearance changes. The proposed method provides a solution to this problem by an adaptive and optimized visual model.

For the situation of occlusion, two videos "faceocc1" and "faceocc2" are employed for test. In the two videos, a woman face is occluded by a book in different directions, and during the occlusion, the appearance of the target and background will take great changes. The quantized evaluation of tracking results is shown in Figure 1. Over the sequence, the face is occluded totally or partially for many times. The proposed methods employing 3-cues and 49-cues are able to handle such a situation. In comparison, the optimized model overwhelms the fixed model and the adaptive model without optimization. When the occlusions happen, it challenges the observation model, which is important in discriminating the target from background. The success of the proposed method lies in the optimized observation model to some extent. In this model, it tries to find a best description of the target appearance. The success also can be seen from the tracking results of Figure 2.

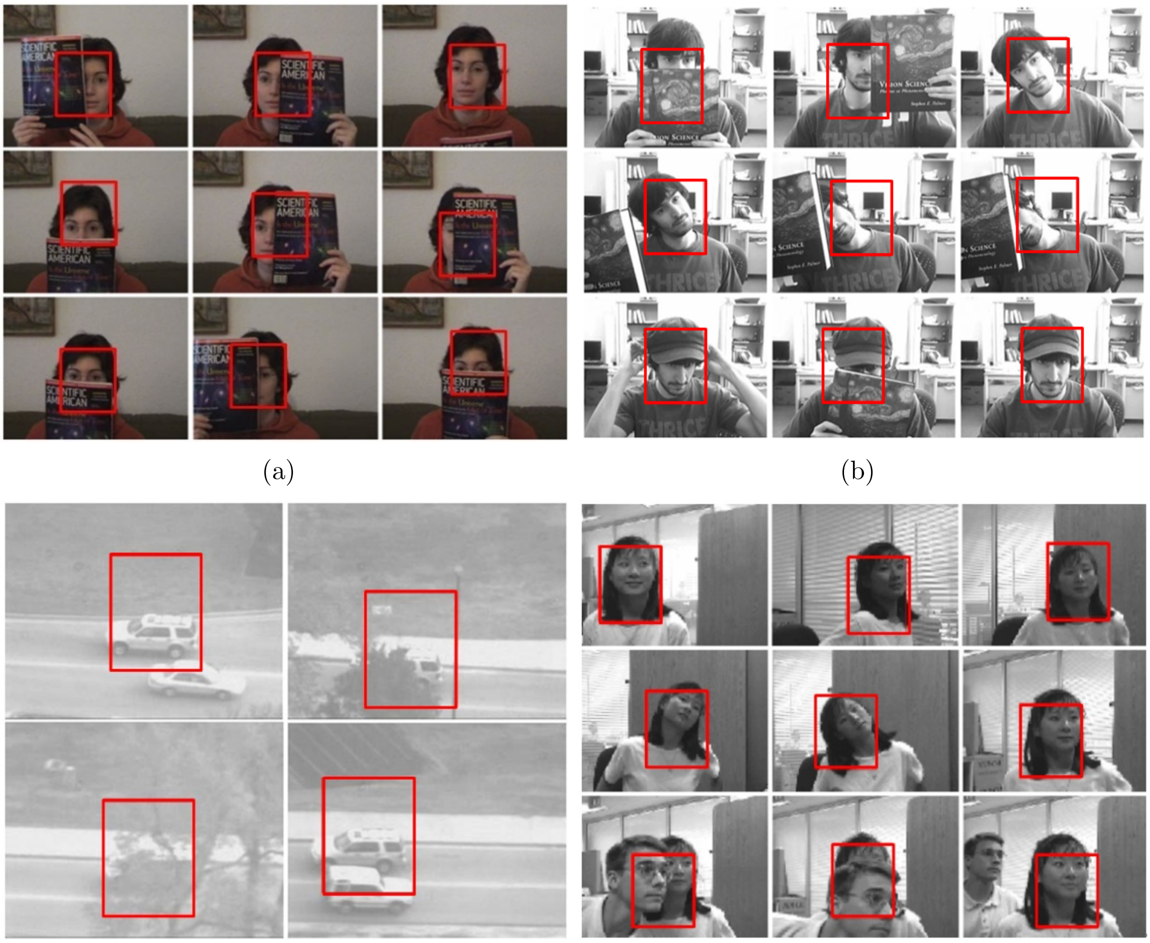


(a)

(b)

Figure 1. The ACLE and AOR curves of test on video "faceocc2". (a) and (b) are the curves tested on 3-cues and 49-cues integration model, respectively. The ACLE values are the smaller the better, and the AOR values are the larger the better.

4.2. **Complex background.** For visual problem, the objects beside of the target are the backgrounds. Various challenges come from the background, such as illumination variation, similar objects, and blurred scene. In order to deal with this kind of situation, the proposed visual model built the visual model based on the analysis of both the target and its background. By this way, the discriminative ability of the built visual model will be better.



(a) (b)

(c) (d)

Figure 2. Tracking results of videos (a) “faceocc1” (frames #103, #180, #245, #306, #457, #556, #62, #703, #874), (b) “faceocc2” (frame #171, #265, #335, #390, #466, #492, #613, #700, #800), (c) “suv” and (d) “girl” (frames #29, #121, #288, #315, #326, #375, #429, #439, #501)

The complex background situation is tested on two challenging videos “suv” and “girl”. For the video “suv”, a car is moving on a road, and over the sequence, there are similar cars passing and trees occlusion. For the video “girl”, a target girl is moving with rotation, occlusion, and illumination changes. In handling such cases, our method realizes robust tracking by constructing a robust appearance model. The tracking results are shown in Figure 2. As can be seen from these tracking results, our method realizes stable tracking in the challenges as blurred scene and similar objects. The success relies on the optimized visual model.

5. **Conclusions.** This paper proposed to resolve the challenges in visual tracking by constructing a robust visual model. The main task of this paper is employing genetic algorithm to optimize the integration model, specifically, the integration weights of all the cues. By defining the fitness function, the visual model is optimized to be adaptive with the tracking condition changes. The robustness and effectiveness are demonstrated by the test on benchmark videos. The selection of cues with best discriminate ability may be a main direction of this work.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant No. 61300099, the China Postdoctoral Science Foundation funded project under Grant No. 2015M570261, and the Science and Technology Development Plan of Jilin Province under Grant No. 20170101144JC, the open fund of China Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education under Grant No. 93K172016K14, and the Fundamental Research Funds for the Central Universities No. 2412017FZ027.

REFERENCES

[1] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick and A. V. D. Hengel, A survey of appearance models in visual object tracking, *ACM Trans. Intelligent Systems and Technology*, vol.4, no.4, pp.1-58, 2013.

[2] L. Ahn, B. Maurer, C. McMillen, D. Abraham and M. Blum, reCAPTCHA: Human-based character recognition via web security measures, *Science*, 2008.

[3] R. Snow, B. O'Connor, D. Jurafsky and A. Ng, Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks, *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2008.

[4] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni and L. Moy, Supervised learning from multiple experts: Whom to trust when everyone lies a bit, *Proc. of International Conference on Machine Learning*, 2009.

[5] H. Wu, A. Sankaranarayanan and R. Chellappa, Online empirical evaluation of tracking algorithms, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2010.

[6] B. Zhong, H. Yao, S. Chen, R. Ji, X. Yuan, S. Liu and W. Gao, Visual tracking via weakly supervised learning from multiple imperfect oracles, *Proc. of Conference on Computer Vision and Pattern Recognition*, 2010.

[7] H. Grabner and H. Bischof, On-line boosting and vision, *Proc. of Conference on Computer Vision and Pattern Recognition*, 2006.

[8] M. Isard and A. Blake, CONDENSATION – Conditional density propagation for visual tracking, *International Journal of Computer Vision*, 1998.

[9] D. Comaniciu, V. Ramesh and P. Meer, Kernel-based object tracking, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2003.

[10] I. Matthews, T. Ishikawa and S. Baker, The template update problem, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2004.

[11] M. Yang, J. Yuan and Y. Wu, Spatial selection for attentional visual tracking, *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[12] H. Grabner, C. Leistner and H. Bischof, Semi-supervised on-line boosting for robust tracking, *Proc. of European Conference on Computer Vision*, 2008.

[13] M. Kolsch and M. Turk, Fast 2D hand tracking with flocks of features and multi-cue integration, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, 2004.

[14] J. Hoey, Tracking using flocks of features, with application to assisted handwashing, *Proc. of British Machine Vision Conference*, pp.367-376, 2006.

[15] Z. Yin and R. Collins, On-the-fly object modeling while tracking, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1-8, 2007.

[16] B. Martinez and X. Binefa, Piecewise affine kernel tracking for non-planar targets, *Pattern Recognition*, vol.41, no.12, pp.3682-3691, 2008.

[17] W. Chang, C. Chen and Y. Hung, Tracking by parts: A Bayesian approach with component collaboration, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol.39, no.2, pp.375-388, 2009.

[18] J. S. Kwon and K. M. Lee, Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping Monte Carlo sampling, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1208-1215, 2009.

[19] L. Cehovin, M. Kristan and A. Leonardis, Robust visual tracking using an adaptive coupled-layer visual model, *IEEE Trans. Pattern Recognition and Machine Intelligence*, 2012.

[20] N. Wang and D. Y. Yeung, Learning a deep compact image representation for visual tracking, *Advances in Neural Information Processing System*, pp.809-817, 2013.

[21] L. Wang, W. Ouyang, X. Wang and H. Lu, Visual tracking with fully convolutional networks, *Proc. of IEEE Conf. Comput. Vis.*, pp.3119-3127, 2015.

[22] C. Ma, J. B. Huang, X. Yang and M. H. Yang, Hierarchical convolutional features for visual tracking, *Proc. of IEEE Conf. Comput. Vis.*, pp.3074-3082, 2015.

[23] H. Nam and B. Han, Learning multi-domain convolutional neural networks for visual tracking, *Computer Science*, 2016.

[24] Z. Cui, S. Xiao, J. Feng and S. Yan, Recurrently target-attending tracking, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp.1449-1458, 2016.

[25] Y. Wu, J. Lim and M.-H. Yang, Object tracking benchmark, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.37, no.9, pp.1834-1848, 2015.

[26] Y. Wang, L. Jiang, Q. Liu and M. Yin, Optimal appearance model for visual tracking, *PLOS ONE*, vol.11, no.1, pp.1-15, 2016.