Proceedings of the 2017 IEEE
International Conference on Information and Automation (ICIA)
Macau SAR, China, July 2017

Fast Tracking via Spatio-Temporal Context Learning based on Multi-color Attributes and PCA*

Yixiu Liu, Yunzhou Zhang, Meiyu Hu, Pengju Si and Chongkun Xia
College of Information Science and Engineering
Northeastern University
Shenyang, China 110819
zhangyunzhou@mail.neu.edu.cn

***Abstract*— At present, the effective tracking of pedestrians is still a challenging task due to factors such as illumination change, pose variation, motion blur and occlusion. In this paper, we propose a simple and effective tracking algorithm which exploits the spatio-temporal context. Based on a existing Bayesian framework, we take full advantage of the relevance of the region of interest to its local context, and model the correlation of the visual characteristics of the target and the surrounding area. Through computing a confidence map and maximizing the likelihood function, our tracker gets the real position of the target. Multi-dimensional variant of color attributes provides superior performance for visual tracking in recent years. Our tracker extracts 11 dimensional color names features from the target. However, considering the real time tracking, we reduce the dimensionality of features from 11 to 3 with PCA algorithm. In order to ensure robustness, we cascade the HSV histogram features. In the absence of optimization, our algorithm runs at more than 80 frames per second implemented in MATLAB. Extensive experimental results show that the proposed algorithm is more accurate and accurate than many state-of-the-art algorithms on multiple datasets.**

***Index Terms*— Spatio-temporal context; color name; HSV histogram; PCA**

I. Introduction

In the field of computer vision, tracking is a fundamental problem, with applications in video surveillance, human-machine interfaces and robot perception. Although a large number of algorithms have been proposed in recent years, target tracking is still a challenging issue due to factors such as pose change, occlusion, illumination, relative movement between targets and so on. All above situations will change the appearances of the targets. Designing a fast and efficient tracker is still extremely hard.

Tracking algorithms can be generally categorized as either generative [1, 2, 3, 4, 5] or discriminative [6, 7, 8, 9] based on their appearance models.

Generative tracking algorithms are used to express the target by establishing an appearance model, and then determine the image region of the target according to the best matching score. Unfortunately, a dramatic target appearance model will bring drift problems. The discriminant model regard tracking problem as a classification issue, the model judged where the decision boundary is between the image blocks and background images. As a result, the accuracy rate is higher than the generative ones. For feature extraction, it may require a large number of samples to train the classifier, which usually takes computing pressure. In a word, which one to be chosen should balance between efficiency and the precision.

In general, target tracking mainly has difficulty in following three aspects. Although the use of color system can reduce the influence of illumination change, it still can?t completely eliminate the effects of illumination changes. Therefore the illumination variation is still an important issue [10]. Occlusion problem, tracking failure due to occlusion by other objects is also a tricky problem in target tracking. Initialization is the target positioning of the first frame in the video. Inaccurate positioning may have great negative impact on the subsequent tracking.

Recently, a fast tracking algorithm that we call it STC algorithm gets the real position of the target through utilizing the spatio-temporal context [11]. However, the visual features it uses are too simple. This approach inspired us to design more robust visual features to make tracking better. We extract 11 dimensional color names features which show great performance in tracking task in recent years. The method combines the principal component analysis (PCA) with the STC algorithm, which takes advantage of temporal and spatial information of the target. In addition to that, we cascade the traditional hand-crafted descriptor, HSV histogram, which has good performance in recognition tasks. Plenty of experiences have been designed to compare our method with several states of art methods. We test our method both on open datasets and real environment, and all of the experiment results have illustrated that our method is effective and valid.

II. Related Works

In this section, we briefly review the relevant works that have been done by other individuals. The differences between

*This work is partly supported by National Natural Science Foundation of China (61471110), Foundation of Liaoning Provincial Department of Education(L2014090), Chinese Universities Scientific Foundation(N160413002, N16261004-2/3/5)

978-1-5386-3154-6/17/$31.00 ©2017 IEEE

other trackers and ours are explained briefly.

Since most trackers either rely on simple colors or use luminance information to represent images, Joost investigated how color names learned from images in real-world rather than the color chips. They proposed several variants of the Probabilistic Latent Semantic Analysis (PLSA) model, which can learn color names from the data with noise. In this way, color names that learned from real-world images significantly outperforming over those learned from color chips that have been labeled for both image retrieval and image annotation [12, 13]. The learned Color Name is robust in color space. Inspired by this, we intend to use color space to increase the robustness of the tracking system. Danelljan presented a method, which makes use of adaptive color attributes, they exploit sophisticated color features that combined with luminance [14]. However, their algorithm structure is too complex, which may lead to a decline in real-time. Conversely, we will use the elegant Bayesian framework in this paper.

Zhang proposed a target tracking algorithm, which is based on the appearance model. The features can be extracted from a multi-scale image feature space, in which the data is independent between each other. Their models exploit non-adaptive random projection matrix, retaining the structure of the target image feature space. Afterwards, a naive binary Bayes classifier with online update is used in the compressed domain [15]. However, updating the classifier in each of the frames will lead to cumulative errors, which will cause drift. Some of the errors are due to the target being partially obscured and some are due to over-samplings of the background samples, which will lead to inaccurate labeling of the target.

The tracking algorithm, which makes use of the information of spatio-temporal context has an outstanding performance in celerity. Through Bayesian formula, they describe the spatial and temporal relationship between the object and its local context. After establishing a statistical relationship of the low-level features of the target and its surrounding area, we can maximize a likelihood function. Thus the most probable position of the target can be obtained, which translates the problem into computing confidence map. Afterwards, the Fast Fourier Transform (FFT) is utilized for fast tracking [11]. The Bayes framework as a simple and elegant method, is used in tracking process has achieved good results and stable effect. However, the method only makes use of low-level features of the target, which can't be adapted for complex scenes. In this way, we plan to improve the method from here. In other words, we use elegant Bayesian framework, but select more robust features.

Chen makes full use of target material, color and other information, the method based on matrix shift, using the Bhattacharyya coefficient to measure the similarity in co-occurrence matrices, realizing human face tracking, unfortunately, their system is too complex for us, which is impractical. He, S. presented locality sensitive histogram algorithm for visual tracking. The local sensitive histogram is calculated at the position of each pixel. For each luminance value, the floating-point value will be added to the corresponding bin [16]. Cognitive psychology is also used to create a robust appearance model that can perform well in both short-term and long-term tracking [17]. Recently, some scholars have proposed a robust algorithm based on multi-tracker integration and multi-feature selection interaction mechanism [18].

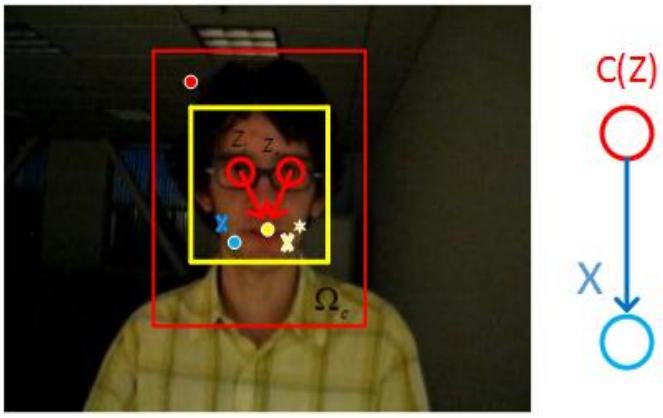


Fig. 1: Graphical model representation of spatial relationship between object and its local context. The context feature at location z is denoted by $c(z) = (I(z), z)$ including appearance representation(i.e.,image feature $I(z)$) and location information.

In conclusion, there is a good prospect in the field of visual object detection and tracking. Therefore, we present a novel method both takes advantage of the elegant Bayes framework and the robust color name attribute. The experiments are implemented to verify the proposed algorithm, which is superior to some other state-of-art methods.

III. PROPOSED TRACKING ALGORITHM

In this section, we present our tracking algorithm in details. consider the spatial correlation of the object in the scene, we obtain the model between the region of interest and its local context. Next, the model get update both in time and space according to the statistical correlation between the visual features from surrounding region and the target. We utilize fast Fourier transform (FFT) to simplify the calculation, so that the tracking process can be carried out in real time. Finally, the tracking problem is transformed into a problem of computing the likelihood of the target position by calculating the confidence pattern:

$$c(x) = P(x|o) \tag{1}$$

Where the target present in the scene is indicated by o, and x is the location of the target. Then we make full use of

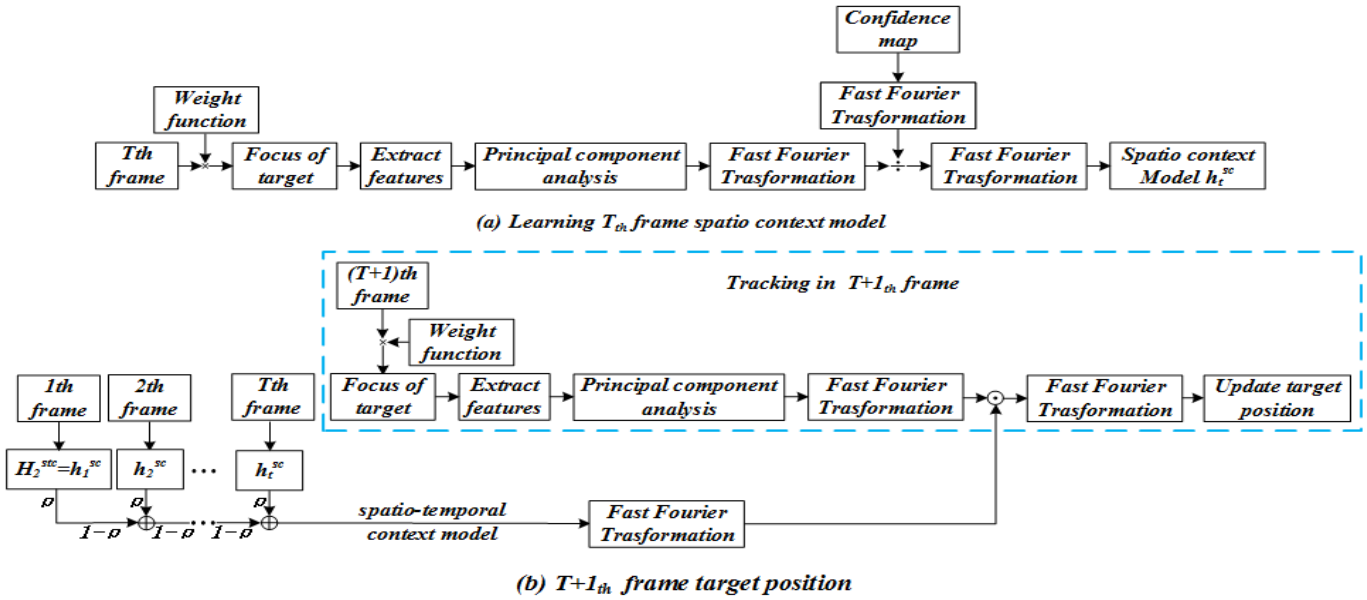


Fig. 2: Main components of our tracking algorithm

spatial context information to estimate the target location. Figure 1 shows its graphical model representation. It is noteworthy that the features we extract from the object and its surrounding background are color names, and we exploit the trick to reduce the dimensional, which makes the process more efficient.

A. spatio-temporal model for tracking

In this section, we mainly formulate the components or overall tracking algorithm as described in Figure 2. We use $P(x|c(z),o)$ to represent the conditional probability function.

$$P(x|c(z),o) = h^{sc}(x-z) \tag{2}$$

We use $h^{sc}(x-z)$ to represent the relationship between the target x and its local context z in distance and direction.

The context prior probability is simply modeled by

$$P(c(z)|o) = I(z)\omega_\sigma(z-x^*) \tag{3}$$

where $I(\cdot)$ is multi-dimensional color attributes which represents appearance of context. We use $\omega_\sigma(\cdot)$ to represent weighted function, and we define it by

$$\omega_\sigma(z) = ke^{-\frac{|z|^2}{\sigma^2}} \tag{4}$$

k is a normalization constant, and makes the value of parameter $P(c(z)|o)$ be between 0 and 1 in (3), so that the definition of probability can be satisfied. We defined the scale parameter as σ.

We set the confidence pattern that represents the target location into

$$\begin{aligned} c(x) &= P(x|o) \\ &= se^{-\left|\frac{x-x^*}{\alpha}\right|^\beta} \\ &= \textstyle\sum_{c(z)\in X^c} P(x,c(z)|o) \\ &= \textstyle\sum_{c(z)\in X^c} P(x|c(z),o)P(c(z)|o) \end{aligned} \tag{5}$$

where α determines the size of the scale. β is a shape factor (See Figure 3). s is a normalization parameter.

Based on a priori model of the context (3), and the existing confidence graph model (5), we can get the spatial context model (2). Putting (2), (3) and (5) together, we can get the likelihood function about the target position

$$\begin{aligned} c(x) &= se^{-\left|\frac{x-x^*}{\alpha}\right|^\beta} \\ &= \textstyle\sum_{z\in\Omega_c(x^*)} h^{sc}(x-z)I(z)\omega_\sigma(z-x^*) \\ &= h^{sc}(x)\otimes(I(x)\omega_\sigma(z-x^*)) \end{aligned} \tag{6}$$

We can utilize the fast Fourier transform (FFT) to convert (6) from the spatial domain to the frequency domain. $\otimes$ denotes the convolution operator which can be used to calculate the convolution quickly. That is,

$$\mathcal{F}(be^{-\left|\frac{x-x^*}{\alpha}\right|^\beta}) = \mathcal{F}(h^{sc}(x)) \odot \mathcal{F}(I(x)\omega_\sigma(x-x^*)) \tag{7}$$

where $\mathcal{F}$ is used to indicates the fast Fourier transform (FFT). $\odot$ indicates the element-wise product. And then, we

have

$$h^{sc}(x) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(be^{-\left|\frac{x-x^*}{\alpha}\right|^{\beta}})}{\mathcal{F}(I(x)\omega_{\sigma}(z-x^*))}\right) \tag{8}$$

$\mathcal{F}^{-1}$ is used to indicates the inverse FFT function.

B. Feature extraction

In this section, we will describe the extracted features in details. Multi-dimensional variant of color attributes provides superior performance for visual tracking in recent years. As an excellent member of them, color names [12] behaves well in many real-world application, such as image retrieval. We hope to continue to explore its potential in object tracking. In this paper, our tracker extracts 11 dimensional color names features from the target. The following picture is processed by image segmentation with color names.

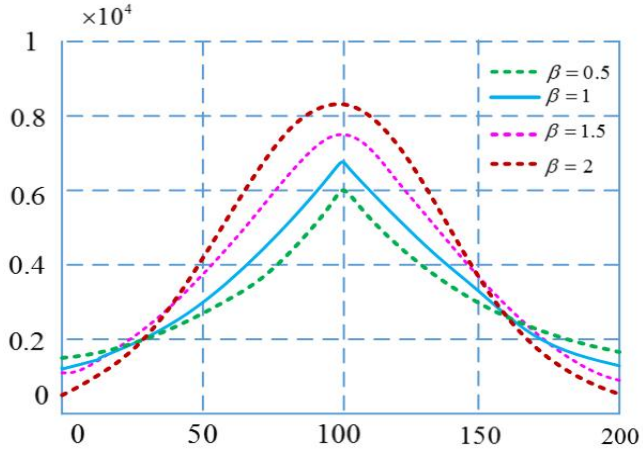


Fig. 3: llustration of $1-D$ cross section of the confidence map $c(x)$ in (5) with different parameters β. Here, the object location $x^* = (100, 100)$.

For the reason of making the visual features more descriptive, we segment the scenes by color names. For every RGB images in the tracking sequence, we label the different parts in black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow colors according to their color attributes [12]. As we can see in the Figure 4, the pedestrians are divided into several parts according to color names. The color names learned from Google Image search, PLSA-ind, obtain also satisfying results for the achromatic regions.

HSV histogram is the traditional hand-crafted descriptor which has good performance in recognition and tracking tasks [19]. In this experiment, we extract $8 \times 8 \times 8$=512 dims HSV histogram features, and cascade them after the color names.

C. PCA for reducing the dimensionality

Tracking must be real-time, otherwise it is meaningless. To make it come true, we exploit PCA algorithm to reduce

Fig. 4: The color names are represented by their corresponding color

redundant information. In our experiment, we reduce the dimensionality of features from 11 to 3 with PCA algorithm. The following is the process of reducing the dimensionality.

Before executing the algorithm, we normalize the image data firstly. Then we have $x^i = x^i - \mu^i$, where x is the feature matrix ($11-D$ color names) extracted from the object, and $x^{(i)}$ indicates the i-th feature vector. If the feature vectors are on different magnitudes, we need to divide it by the standard deviation σ.

Next we calculate the covariance matrix,

$$\sum = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^{\mathrm{T}} \tag{9}$$

where $\sum$ is the covariance matrix, m is the number of dimensions and $x^{(i)}$ is the feature vector after normalization. Through singular value decomposition, we have

$$[U, S, V] = SVD(\sum) \tag{10}$$

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \dots & u^{(n)} \\ | & | & & | \end{bmatrix} \in R^{n\times n} \tag{11}$$

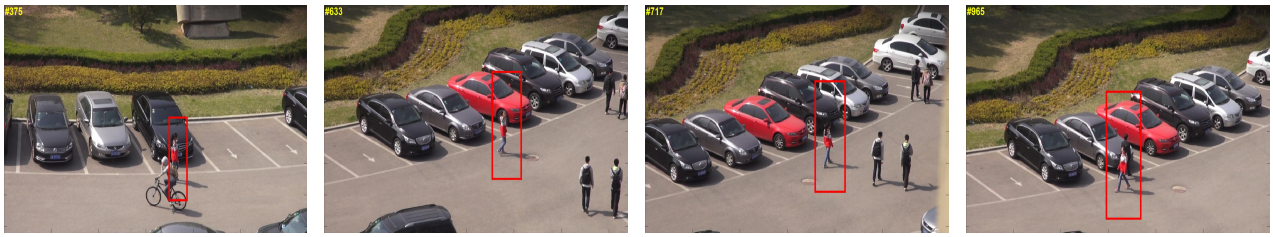
Using the formulas given above, we obtain the projection matrix U. Now we reduce the dimensions of the feature matrix x from $m \times n$ to $n \times k$.

$$z = U_{reduce}^{\mathrm{T}} \times x \tag{12}$$

Finally, we get the features z after reducing the redundant information. In our experience, m is 11, while k is 3. In addition, n is length of the color names vector which is reshaped by image processing toolbox in MATLAB.

(a) (b) (c) (d)

Fig. 5: STC tracker performs in the real scenes



(a) (b) (c) (d)

Fig. 6: Our tracker performs in the real scenes

IV. Experimental result and analysis

In this paper, we show that our tracker performs favorably against many state-of-the-art algorithms on challenging sequences in terms of efficiency, accuracy and robustness. It is worth noting that our tracker is implemented in MATLAB, which runs more than 100 frames per second(FPS). Figure 5 and Figure 6 show that our algorithm behaves better than the original algorithm in real scenes.

A. Experiment results

We evaluate our tracking algorithm with five state-of-the-art methods on six challenging sequences both on success rate(SR)(%) and center location error(CLE)(in pixels). Since most of the trackers involve randomness, we run them 5 times and report the average result for each video clip.

As we can see from Table 1, 2. We experimented with multiple datasets, Compared with other algorithms, our tracker is outstanding in accuracy. We prove that our tracker is really reliable in practical application.

TABLE I: Sucess rate(SR)(%). Comparison with the other five algorithms. our tracker is outstanding in accuracy.

datasets	STC	TLD	struct	CT	VTD	Ours
snowboard	91	80	71	76	38	92
animal	76	37	80	60	79	77
faceocc	82	80	60	70	70	87
faceocc2	84	83	67	66	73	89
02_david	84	81	33	77	82	82
girl	87	79	85	63	68	90

TABLE II: Center location error(CLE)(in pixels) and average frame per second(FPS).

datasets	STC	TLD	struct	CT	VTD	Ours
snowboard	8	13	20	18	40	7
animal	14	125	19	30	16	18
faceocc	10	14	15	28	78	9
faceocc2	12	10	10	30	100	8
02_david	13	13	63	14	14	15
girl	15	16	11	25	119	11

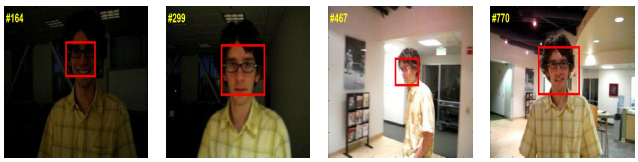
B. Experiment on datasets

Our algorithm is an improvement version for STC algorithm. We extend the color attribute on the basis of the original version. So that the new tracker performs against it on many challenging sequences in accuracy and robustness. However, considering the real-time tracking, we reduce the dimensionality of the color attribute from 11 to 3. In order to ensure robustness, we cascade the HSV histogram features. In the following, we prove that the method is a fast yet robust algorithm in many sequences and real scenes.

Focusing on illumination changes (Figure 7), poses changes, (Figure 8) and partly occluded problems (Figure 9), we experimented in the corresponding scenes. Finally, all of the experimental results show that our tracker has good performance in dealing with these questions.

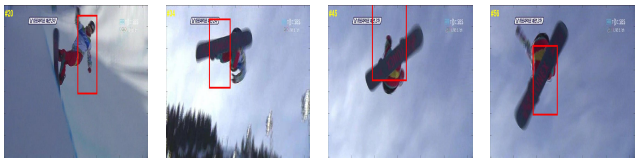
V. Conclusion

In this paper, we present an improved version of a tracking system called STC Track that exploit the spatio-temporal relationships between the object and its local context. It's a pity that the features which the STC Track extracts are



(a) (b) (c) (d)

Fig. 7: Our tracker performs in the scenes where illumination changes obviously



(a) (b) (c) (d)

Fig. 8: Our tracker performs in the scenes where the target's poses change obviously

(a) (b) (c) (d)

Fig. 9: Our tracker performs in the scenes where the target is partly occluded

the image intensity and its position. It lead us to extend the multi-color attributes on the basis of the original version. To make sure the tracker performs in real time, we reduce the redundant information while extracting features.

There are many interesting ways to extend this work in the future. Firstly, the multi-color attributes can be replaced by the deep features which are recognized as the better features if the process of training is effective. Furthermore, to select more useful features, some methods can work better such as sparse coding. Finally, we are interested in other possible application for our algorithm.

REFERENCES

[1] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[2] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1296–1311, Oct 2003.

[3] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vision*, vol. 77, no. 1-3, pp. 125–141, May 2008.

[4] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, Nov 2011.

[5] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *CVPR 2011*, June 2011, pp. 1305–1312.

[6] S. Avidan, "Support vector tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1064–1072, Aug 2004.

[7] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *British Machine Vision Conference 2006, Edinburgh, Uk, September*, 2006, pp. 47–56.

[8] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *European Conference on Computer Vision*, 2008, pp. 234–247.

[9] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, Aug 2011.

[10] M. Shah, O. Javed, and K. Shafique, "Automated visual surveillance in realistic scenarios," *IEEE MultiMedia*, vol. 14, no. 1, pp. 30–39, Jan 2007.

[11] K. Zhang, L. Zhang, M. H. Yang, and D. Zhang, "Fast tracking via spatio-temporal context learning," *Computer Science*, 2013.

[12] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, July 2009.

[13] J. van de Weijer, C. Schmid, and J. Verbeek, "Learning color names from real-world images," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[14] M. Danelljan, F. S. Khan, M. Felsberg, and J. v. d. Weijer, "Adaptive color attributes for real-time visual tracking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1090–1097.

[15] K. Zhang, L. Zhang, and M. H. Yang, "Fast compressive tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 10, pp. 2002–2015, 2014.

[16] S. He, Q. Yang, R. W. H. Lau, J. Wang, and M. H. Yang, "Visual tracking via locality sensitive histograms," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2427–2434.

[17] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 749–758.

[18] J. H. Yoon, M. H. Yang, and K. J. Yoon, "Interacting multiview tracker," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 903–917, May 2016.

[19] C. Xiao, W. Chen, and H. Gao, "Object tracking algorithm based on hsv color histogram and block-sparse representation," in *2015 34th Chinese Control Conference (CCC)*, July 2015, pp. 3826–3831.