2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

A Benchmark for Full Rotation Head Tracking

Yulin Li*†, Bingpeng Ma*†, Hong Chong*, Xilin Chen*†
*Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS),
Institute of Computing Technology, CAS, Bejing, China
†School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China
Email: {yulin.li, bingpeng.ma, hong.chong, xilin.chen}@vipl.ict.ac.cn

***Abstract*—This paper introduces a new benchmark for 360-degree rotation head tracking, named Full Rotation Head Tracking (FRHT). The benchmark consists of 50 color sequences containing diverse human activities with complicated head motions. Specially, FRHT covers the most challenges of head tracking and focuses on the appearance variations of heads during the 360-degree rotation. It also pays attention to the clutters from the heads of nearby people. Further, we propose a baseline tracker. It guides a selective adaption updating by verifying strategies, thus alleviates error accumulation. Extensive experiments validate the advantages of FRHT in head rotation and similar object clutter.**

I. Introduction

Head tracking, estimating serial locations of a certain human head in a video, has found wide applications in computer vision, including vehicle navigation, human-computer interaction and surveillance. During the past years, several works [2,3,10,21,24] have been proposed to improve the head tracking performance. However, this task is far from settled due to miscellaneous difficulties.

The main challenge of head tracking comes from a great appearance change caused by head rotation. For example, when the person turns around, the back of his head instead of the face, becomes visible. Most head trackers suffer from this large degree rotation thus fail to track. To tackle this issue, head rotation need gain more attention.

However, the absence of proper benchmark has seriously hindered the progress of head tracking. To the best of authors' knowledge, only one benchmark [2] has been published for full 360-degree rotation during the past two decades. Though general tracking benchmarks [13,17,20] include a subset of head sequences, which mainly focus on faces with slightly changing. The missing of rotations results in limited research and application for head tracking. So a complete benchmark for head tracking is urgently demanded.

In this paper, we first establish a benchmark named Full Rotation Head Tracking (FRHT) to evaluate head tracking with full rotations in real-world scenes. The benchmark takes primarily care of the 360-degree viewpoint and head rotation. Meanwhile, it emphasizes a more ubiquitous challenge of similar object clutters in head tracking. To our best knowledge, FRHT is the largest benchmark to exhibit the challenges and evaluate algorithm performance towards head tracking.

In addition, we design a novel head tracker with verification (HTV) to address the challenges in FRHT. More specifically, we introduce a verification framework, where the head specific knowledge can be utilized for building a selective updating strategy. In this way, the accumulative errors caused by online updating are greatly alleviated. Extensive evaluations demonstrate the effectiveness of our benchmark and algorithm.

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

II. Related Work

In this section, we briefly review the related work, from the perspectives of the benchmark and algorithm.

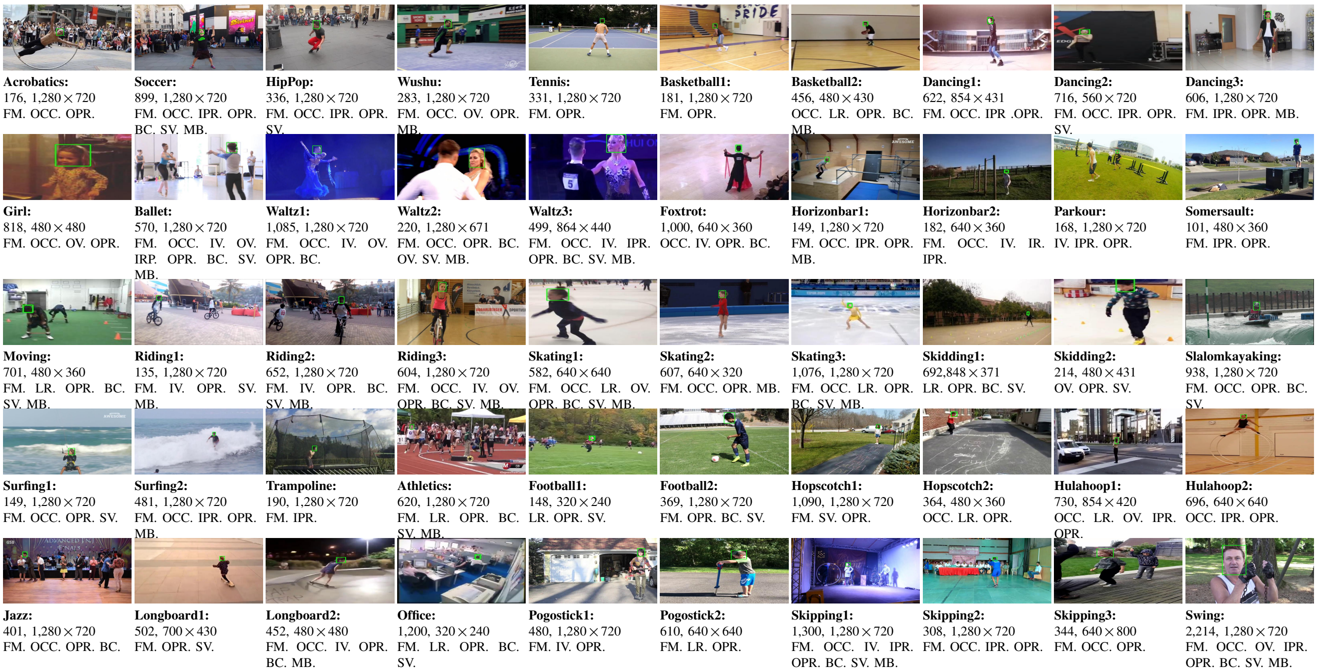
Benchmark. In 1998, Birchfield [2] proposed the first benchmark for full 360-degree rotation head tracking named Headtracker. It is captured under the laboratory environment and contains 16 short videos, totally 2,166 frames. 14 videos of them have elliptical head annotations and contain one to three disturbing conditions. Considering its data size and diversity of scenarios, the benchmark is limited to promote the development of head tracking algorithms.

Fortunately, several benchmarks, such as OTB100 [20], VOT [13] and ALOV300++ [17], include head sequences. For instance, the head subset of OTB100, we called OTB100 Head, consists of 23 sequences with 11,307 frames. Diverse scenes OTB100 Head includes are challenging to modern trackers. However, the benchmark lacks large degree head rotations. It is heavily weighted towards faces and produces an extremely imbalanced distribution of head viewpoints. As a result, it is biased to perform evaluations for head tracking algorithms.

In this paper, we develop a new head tracking benchmark. Compared with the above benchmarks, It cares more about full 360-degree head rotation in complex circumstances.

Algorithms. Most recent studies [2,3,10,21,24] focus on the design of head appearance model to improve the robustness to rotation. Li *et al.* [10] proposed a multi-state particle filter for inferring face and head states by two particle groups. Kiyotake *et al.* [21] related the appearance model with the camera parameters, to adaptively select the model for different head views. Bouaynaya *et al.* [3] incorporated motion estimation into mean shift algorithm to generate trajectories of targets.

In addition, the combination of multiple cues is studied for head tracking. Birchfield [2] introduced color histogram and intensity gradient into an elliptical model for reliable tracking. In order to improve the discrimination of targets, Zhang *et al.* [24] proposed a kernel Bayesian framework based on multiple cues. The major drawback of above algorithms lies in the hand-crafted features which are limited in handling the significant appearance caused by head rotation. Recently, many



Acrobatics: 176, 1,280×720 FM. OCC. OPR.
Soccer: 899, 1,280×720 FM. OCC. IPR. OPR. BC. SV. MB.
HipPop: 336, 1,280×720 FM. OCC. IPR. OPR. SV.
Wushu: 283, 1,280×720 FM. OCC. OV. OPR. MB.
Tennis: 331, 1,280×720 FM. OPR.
Basketball1: 181, 1,280×720 FM. OPR.
Basketball2: 456, 480×430 OCC. LR. OPR. BC. MB.
Dancing1: 622, 854×431 FM. OCC. IPR .OPR.
Dancing2: 716, 560×720 FM. OCC. IPR. OPR. SV.
Dancing3: 606, 1,280×720 FM. IPR. OPR. MB.
Girl: 818, 480×480 FM. OCC. OV. OPR.
Ballet: 570, 1,280×720 FM. OCC. IV. OV. IRP. OPR. BC. SV. MB.
Waltz1: 1,085, 1,280×720 FM. OCC. IV. OV. OPR. BC.
Waltz2: 220, 1,280×671 FM. OCC. OPR. BC. OV. SV. MB.
Waltz3: 499, 864×440 FM. OCC. IV. IPR. OPR. BC. SV. MB.
Foxtrot: 1,000, 640×360 OCC. IV. OPR. BC.
Horizonbar1: 149, 1,280×720 FM. OCC. IPR. OPR. MB.
Horizonbar2: 182, 640×360 FM. OCC. IV. IR. IPR.
Parkour: 168, 1,280×720 IV. IPR. OPR.
Somersault: 101, 480×360 FM. IPR. OPR.
Moving: 701, 480×360 FM. LR. OPR. BC. SV. MB.
Riding1: 135, 1,280×720 FM. IV. OPR. SV. MB.
Riding2: 652, 1,280×720 FM. IV. OPR. BC. SV. MB.
Riding3: 604, 1,280×720 FM. OCC. IV. OV. OPR. BC. SV. MB.
Skating1: 582, 640×640 FM. OCC. LR. OV. OPR. BC. SV. MB.
Skating2: 607, 640×320 FM. OCC. OPR. MB.
Skating3: 1,076, 1,280×720 FM. OCC. LR. OPR. BC. SV. MB.
Skidding1: 692,848×371 LR. OPR. BC. SV.
Skidding2: 214, 480×431 OV. OPR. SV.
Slalomkayaking: 938, 1,280×720 FM. OCC. OPR. BC. SV.
Surfing1: 149, 1,280×720 FM. OCC. OPR. SV.
Surfing2: 481, 1,280×720 FM. OCC. IPR. OPR. MB.
Trampoline: 190, 1,280×720 FM. IPR.
Athletics: 620, 1,280×720 FM. LR. OPR. BC. SV. MB.
Football1: 148, 320×240 LR. OPR. SV.
Football2: 369, 1,280×720 FM. OPR. BC. SV.
Hopscotch1: 1,090, 1,280×720 FM. SV. OPR.
Hopscotch2: 364, 480×360 OCC. LR. OPR.
Hulahoop1: 730, 854×420 OCC. LR. OV. IPR. OPR.
Hulahoop2: 696, 640×640 OCC. IPR. OPR.
Jazz: 401, 1,280×720 FM. OCC. OPR. BC.
Longboard1: 502, 700×430 FM. OPR. SV.
Longboard2: 452, 480×480 FM. OCC. IV. OPR. BC. MB.
Office: 1,200, 320×240 FM. LR. OPR. BC. SV.
Pogostick1: 480, 1,280×720 FM. IV. OPR.
Pogostick2: 610, 640×640 FM. LR. OPR.
Skipping1: 1,300, 1,280×720 FM. OCC. IV. IPR. OPR. BC. SV. MB.
Skipping2: 308, 1,280×720 FM. OCC. IPR. OPR.
Skipping3: 344, 640×800 FM. OCC. OPR.
Swing: 2,214, 1,280×720 FM. OCC. OV. IPR. OPR. BC. SV. MB.

Fig. 1. The first frame of each FRHT sequence is shown with the green bounding box of the target head. The first row shows the sequence name and the three items of the rest rows illustrate the frame length, resolution and the attributes defined in the Table II, respectively. Some pictures are resized for display.

deep learning based trackers [1,5]–[7,12] take full advantages of Convolutional Neural Networks (CNN) to obtain semantic features and achieve excellent performance in visual tracking.

Here, we propose a new tracking method based on verification framework. Compared with aforementioned methods, our tracker integrates the head prior knowledge into the online updating and finally relieves the tracking drift.

III. The FRHT Dataset

In this section, we introduce our new FRHT benchmark. Fig. 1 shows the first frames of all the 50 sequences of FRHT with bounding box annotations.

A. Annotation protocol

For each target, the annotation is manually tagged by a rectangular bounding box with temporal smoothness. For each sequence, several challenge attributes are also annotated. All the annotations are obtained according to the following instructions: (1) All head poses from the front to the back are annotated. (2) For the front, the bounding box goes vertically from the chin to the forehead, and horizontally from one ear to the other or the nose depending on the pose. (3) For the backend, the bottom line of the bounding box is demarcated at the top of the neck. (4) Particularly, in the case that the ear and neck are not completely visible due to occlusion, we determine the bounding box according to the relative position of the eyes or shoulders.

Note that we do not involve the out-of-view portion into the annotated bounding box. In rare cases, when the target moves very rapidly, the annotation is acquired based on the linear interpolation of the temporally adjacent bounding boxes.

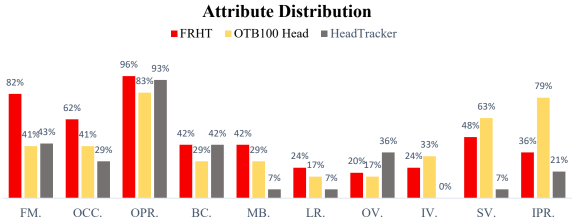


Fig. 2. A comparison of attributes between FRHT, OTB100 Head and HeadTracker.

B. Challenge attribute

FRHT expresses the diverse head movements in real-world conditions. The benchmark videos are captured from the Internet and with a wide variety of scenes (e.g. street, sea, gymnasium, stage, grassland, ice rink) and activities (e.g. running, cycling, surfing, dancing, skating, flying). Naturally, it covers the most challenges of tracking. Specially, we define 10 attributes according to [20], including illumination variation (IV), scale variation (SV), occlusion (OCC), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), and low resolution (LR). Table I gives a detailed introduction for each attribute factor. In Fig. 1 and Fig. 2, we show the challenging attributes of each sequence and total attribute distributions in FRHT, respectively.

C. Detail analysis

FRHT has some advantages over previous related object tracking benchmarks. In the following, we analyze FRHT from the properties of viewpoint, rotation and similar object clutter.

TABLE I
ATTRIBUTES ANNOATED TO TRACKING SEQUENCES.

Attr	Description
FM	Fast Motion: the motion of the bounding box is larger than 20 pixels between adjacent frames.
OCC	Occlusion: the target is partially or fully occluded.
IV	Illumination Variation: the illumination of the target changes significantly
LR	Low Resolution: the ground truth contains less than 400 pixels bounding box.
OV	Out-of-View: some portion of the target leaves the view.
IPR	In-Plane Rotation: the target rotates in the image plane.
OPR	Out-of-Plane Rotation: the target rotates out of the image plane.
BC	Background Clutter: the background near the target has similar appearance as the target.
SV	Scale Variation: the ratio of current bounding box is outside the range [0.5, 2] of the initial one.
MB	Motion Blur: the target is blurred due to the fast motion.

TABLE II
A COMPARISON BETWEEN HEAD TRACKING DATASETS IN THE LITERATURE. THE FIRST RANK IS HIGHLIGHTED IN RED.

Datasets	Headtracker	OTB100 Head	FRHT
Scenes	single indoor	mixture	mixture
Full Annotation	no	yes	yes
Number of Seq.	16	23	50
Min Frames	31	100	148
Max Frames	500	1,490	2,214
Total Frames	2,166	11,307	28,247
Min Target Size (pixels)	1,080	120	85
Max Target Size (pixels)	3,000	50,096	57,096
Number of Rotation	11	4	285
Speed of Rotation	40	-	15
Similar Object Clutter	few	none	many

Viewpoint. According to the variations of camera views, each target head has multiple viewpoints. An imbalanced distribution of viewpoints will lead to the biased evaluation of tracking algorithms. FRHT has a relatively uniform distribution of viewpoints. By observing the difference of appearances, we separate the head viewpoints into the front, profile and backend. The proportions of the three parts in FRHT are 35%, 38% and 27%, respectively.

For comparison, we also evaluate the viewpoints of OTB100 Head. The statistical result shows that only 4 sequences display sort of the backend of heads while the remaining 19 sequences focus on faces. Therefore, in this regard, our proposed benchmark provides complete head appearances, which is more close to real scenarios.

Head rotation. The large degree head rotation is a great challenge for head trackers. In FRHT, half of the sequences contain 360-degree rotations at least 3 times, and the maximum is 24 times in the sequence of Ballet. Moreover, there are four out-of-plane rotations (left, right, up and down) and two in-plane rotations (clockwise and counter-clockwise). Specially, the proportions of turning left (40%) and right (44%) occupy the substantial portion. The rotations of turning up (7%) and bottom (9%) also appear in some jump or somersault activities.

Contrary to FRHT, existing benchmarks neglect head rotations to a great extend. We give a comparison for FRHT, OTB100 Head and Headtracker in the bottom of Table II. On one side, the rotation frequency of FRHT is far more than other benchmarks'. Our benchmark includes 285 times of full rotations among all the sequences. But OTB100 Head and Headtracker merely have 4 and 11 times, respectively. On the other side, the rotating speed in FRHT is more rapid than that of Headtracker (15 frames per rotation in FRHT versus 40 in Headtracker). These statistics demonstrate that FRHT captures the complicated head rotations, thus is more appropriate for training and evaluating real-world head trackers.

Similar object clutter. Besides rotation, background with similar object clutter is another hard issue for trackers. For instance, the heads without green bounding boxes in the figure 1 (such as Waltz1, Waltz2, Waltz3, Foxtrot and Jazz) are the extra objects of background clutters. Similar object clutter leads that the heads of other people generally create a terrible confusion and disturb the determination of trackers. And in general, people often make activities together, which causes frequent similar object clutters in a crowd.

In this paper, we argue that similar object clutter should be considered in the benchmark of head tracking. Statistically, our benchmark involves 11 sequences (such as Skating, Ballet and Athletics) of similar object clutters. Taking into account this realistic factor, FRHT can develop a better tracking evaluation. Contrarily, OTB100 Head and Headtracker scarcely notice this disturbance.

Other attributes. Besides above three aspects, our benchmark possesses apparent advantages in other challenge attributes of visual tracking. We provide the distribution of 10 attributes related to OTB100 Head and Headtracker in Fig. 2. As is shown in the figure, FRHT is inclined in fast motion, occlusion and out-plane rotation with over 50% coverage. By contrast, FRHT contains more proportions than the other two benchmarks with respect to more than half attributes. The histogram demonstrates that FRHT poses more difficulty from the perspective of visual tracking challenges.

Meanwhile, we give statistics of the three benchmarks in the upper of Table II. Our benchmark contains 28,247 frames, which is the double size of OTB100 Head. Moreover, in FRHT, the pixel numbers of head bounding boxes range from 85 to 57,096. The wide range of target sizes increases the challenge of tracking in scale variation. In addition, the maximum sequence length in FRHT is 2,214 frames. Compared with the other two benchmarks, FRHT gives more benefit for long-term tracking.

IV. PROPOSED ALGORITHM

In this section, the architecture of HTV is introduced. The proposed method adopts the correlation filter-based CNN tracker [16,18] as its basis and draws head verifications into the online updating. Fig. 3 depicts the pipeline of HTV.

A. CFNet

CFNet is a two-stream siamese convolutional network, with shared the parameters to solve a correlation filter h on pairs of target patches. This network locates the target in the candidate image x with the learned filter h:

$$F(x, h) = \varphi(x) \star h, \tag{1}$$

where $\varphi(\cdot)$ yields CNN features, $\star$ expresses the correlation operation calculated in the Fourier domain. F is the estimated correlation response map. Further details about the correlation filter are available in the literature [4]. With this framework, we train CFNet with head sequences to bring a significant improvement in head tracking.

From [16], a simple average moving stragety is used to update h which gives a little benefit. Thus, in our framework, we design new model updating scheme and replace the template feature $\varphi(x)$ with the online version:

$$F_{t+1}(x, h) = \varphi(x_{t+1}) \star h_t, \tag{2}$$

B. Updating verifier

Model updating is a fundamental component in visual tracking. To manage the varying appearance during the whole tracking process, the model is required to update constantly. Most trackers make the assumption that the predictions of trackers keep high accuracy in online updating stage. However, there is an inevitable deviation between the prediction and ground truth, which easily leads to background clutter. In addition, without measuring the degree of deviation in tracking outputs, regular updating easily generates error accumulation and leads to a drift. In this paper, we propose a new verifying framework in online updating to address the above problem. The proposed HTV method verifies the prediction of the tracker and improves the validity of model updating. By this way, the wrong updates can be reduced greatly.

As for human head, we apply the color and reappearance characteristics into the design of verifiers. Formally, given the latest template feature $\varphi_t(x)$, the updating scheme at the next frame uses a selective adaptation:

$$\begin{aligned} \varphi_{t+1}(x) = \quad & \omega_1 \alpha(x_{t+1}) \varphi(x_{t+1}) + \omega_2 \beta(x_1) \varphi(x_1) \\ & + (1 - \omega_1 \alpha(x_{t+1}) - \omega_2 \beta(x_1)) \varphi_t(x), \end{aligned} \tag{3}$$

where α, β are two binary functions for verifications. ω_1 and ω_2 are the learning rates.

Color verifier. Several cues like contour, color and rigidity are commonly used in head tracking. In comparison with contour, color is fairly stable to capture the pattern of head and invariant to occlusion, blur and deformation. So, we construct the first verifier α based on the color prior.

Overall, the verifying process consists of two parts. Firstly, to obtain the color of skin and hair, Gaussian Mixture Model (GMM) is employed to the candidate image x. Then, we can estimate three clusters (skin, hair and background) and their mean vectors m_s, m_h and m_b, respectively. Besides, in order to ensure that the clusters correctly cover the color pattern of head, two constraints should be satisfied: (1) Each color region can be separated from the other. (2) The proportions of skin and hair are dominant. When the constraints are violated, an average moving update is deployed instead.

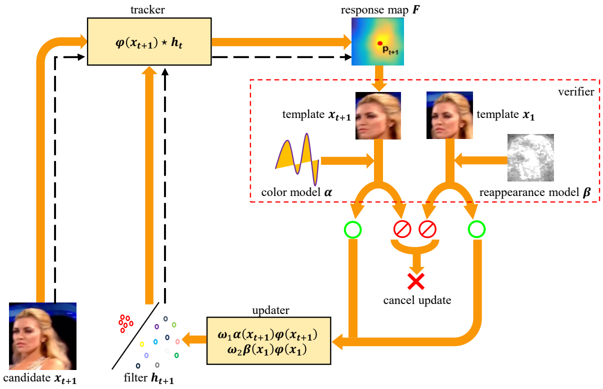


Fig. 3. A basic pipeline of our HTV architecture (orange line). When the tracker (black dashed line) locates and outputs current target observation x_{t+1}, template model $\varphi_t(x)$ is selective updated according to successes of verifiers. Otherwise, the updater is disabled.

Let R^s_{t+1} and R^h_{t+1} represent the pixel sets of skin region and hair region in x_{t+1}, respectively. Based on the GMM, each set consists of pixels whose RGB values are within some specified euclidean distance k_1 from the mean vector of the corresponding GMM component, i.e.,

$$R^*_{t+1} = \{p \mid p \in x_{t+1}, \|p - m_*\|_2 \le k_1\}, * = s \ or \ h. \tag{4}$$

Then, x_{t+1} is verified with α as expressed in Eqn. 5:

$$\alpha(x_{t+1}) = \mathbb{1}\{|R^s_{t+1}| + |R^h_{t+1}| \ge k_2\}, \tag{5}$$

Here $\mathbb{1}\{\cdot\}$ is the boolean indicator function used to verify that the total number of pixels in R^s_t and R^h_t is larger than a given threshold k_2.

Reappearance verifier. Moreover, history information in the tracking process has usually been considered [8,23]. Generally speaking, verification can correct the past tracking mistakes since the view of the head may reappear over time. In consideration of that, the tracking outputs can be gathered as the training data to fine-tune the tracker. Nevertheless, this strategy brings more inaccuracy in tracking outputs as well as higher computational burdens. To avoid slowing down the tracking speed, we select the first template feature $\varphi(x_1)$ from the ground truth to verify the reappearing head.

Specially, when the target template x_1 is reappearing at current frame, there is a higher response than other results. In order to eliminate the numerical difference among sequences, β is activated if the ratio of the peak score of $F_1(x, h)$ to that of $F_t(x, h)$ is greater than a given threshold k_3.

$$\beta(x_1) = \mathbb{1}\{\frac{\max(F_1(x, h))}{\max(F_t(x, h))} \ge k_3\}, \tag{6}$$

V. Experiments

In this section, we evaluate several state-of-the-art trackers on our benchmark. The analysis is given to explain the challenges existing in FRHT and the effectiveness of HTV.

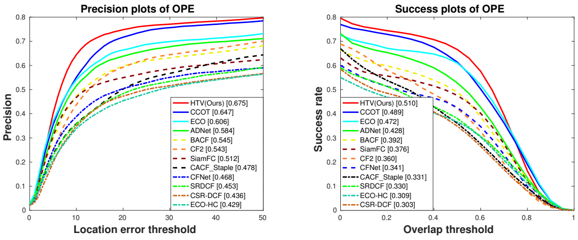


Fig. 4. Overall performance comparison on FRHT.

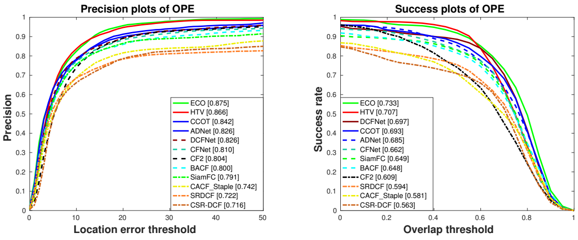


Fig. 5. Overall performance comparison on OTB100 Head.

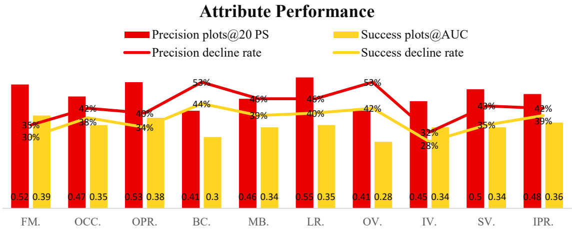


Fig. 6. The average performance of 12 tracks on 10 attributes in FRHT. The bars show the average performance and the lines show the extent of decline.

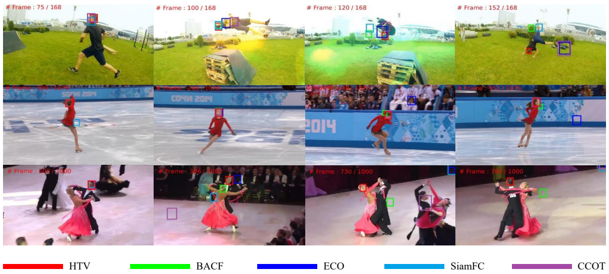


Fig. 7. Qualitative comparison on Parkour, Skating2 and Foxtrot.

A. Implementation detail

The CNN φ is constructed as a 6-layer symmetric encoder-decoder network with 8× down-sampling and up-sampling. The kernel size for convolution and deconvolution is set to 3×3 and the channels are 32,64,128,128,64,32 respectively. Input x is cropped with 2× padding size and resized to 125×125. The training data comes from the head detection dataset HollywoodHead [19]. The model is trained by Adam for 10k iterations with $1e^{-5}$ learning rate. We set $k_1 = 0.6$, $k_2 = 0.6$ and $k_3 = 0.9$. ω_1 and ω_2 are fixed to 0.01 and 0.4. The HTV is implemented in Python with Pytorch [15]. The benchmark and code are available at: *http://vipl.ict.ac.cn/view_database.php*.

B. Benchmark evaluation

Metrics. The traditional metric, Root Mean Square Error (RMSE), computes the standard deviation of all the distances between the tracking results and ground truths. In the case of tracker drift, the metric would become very large and cover up the previous accuracy. Thus, we define that the tracking performance is measured by employing a one-pass evaluation (OPE) based on two metrics: precision plots and success plots [20]. Precision plots show the percentage of frames where the tracking center location stays within a certain distance from the ground truth location. Success plots show the percentage of frames in which overlap ratio between the tracked and ground truth bounding boxes exceeds a threshold.

Quantitative evaluation. On our proposed benchmark and OTB100 Head, we evaluate HTV and other 11 state-of-the-art general trackers including ADNet [22], BACF [9], ECO [5], CACF_Staple [14], CCOT [7], CFNet [18], CF2 [12], CSR-DCF [11], DCFNet [16], SiamFC [1] and SRDCF [6].

Fig. 4 and Fig. 5 illustrate the overall performance on FRHT and OTB100 Head, respectively. As shown in the figures, all the trackers perform remarkably worse on FRHT than on OTB100 Head. For example, ECO, the 2017 tracking champion on OTB100, achieves a high score on OTB100 Head but drops a lot on FRHT. Statistically, the performance of all the trackers decreases by 40% on average with respect to FRHT, which exhibits that our benchmark provides more evident difficulties to these trackers.

For analyzing the performance degradation, Fig. 6 shows the average performance of 12 trackers on each attribute subset. The bars measure the average precision scores and AUCs of success plots. The lines report the performance decline rate from OTB100 Head to FRHT. From the figure, we notice that: Firstly, the performances decrease on all the attributes, which shows the great challenges of FRHT on a finer granularity. Secondly, background clutter (BC) has lower performance (0.41 at precision score and 0.3 at AUC) and higher decline rate (53% and 44%), reflects the strong distraction of similar object clutter from the side. So, this can be seen as the key reason for tracking failure. Finally, out-of-view (OV) is another main distractor. It seems to be hard that trackers catch the target when it gets out of the view and reappears again.

We further show the qualitative results on three challenging sequences. As shown in Fig. 7, BACF, ECO, SiamFC and CCOT have drifts when the target turns around. These cases display the challenge of head rotation in FRHT. Our HTV exhibits a superior ability to handle this condition. Nevertheless, the failure case in Foxtrot shows HTV drifts to the head of the nearby woman. It intuitively depicts the large disturbance of similar object clutter to these tracking methods.

C. Analysis of HTV

We give the results in Fig. 4 and Fig. 5. HTV contributes to a top-tier performance over the two benchmarks. Specially, it outperforms all the trackers on FRHT, as well as achieves

TABLE III
VARIANTS OF HTV FOR ABLATION EVALUATION ON FRHT.

	HTV-N	HTV-G	HTV-β	HTV-α	HTV
α verifier	–	–	–	✓	✓
β verifier	–	–	✓	–	✓
Strategy [6,12]	–	✓	–	–	–
Precision@20 PS	0.605	0.616	0.657	0.660	**0.675**
Success@AUC	0.444	0.447	0.486	0.496	**0.510**

TABLE IV
COMPARISON WITH ZHANG *et al.* 'S METHOD [24] ON HEADTRACKER.

Data	Length(frames)	RMSE	
		Zhang	HTV
Sequence 1	51	3.3879	1.5926
Sequence 4	462	11.6514	5.5707
Sequence 6	310	4.7392	3.1815
Sequence 7	51	2.7395	2.7286
Sequence 8	31	2.3056	2.0959
Average	-	4.9647	3.0339

the second rank on OTB100 Head. These results validate the effectiveness of HTV. In other words, by checking the tracker outputs and employing verified bounding boxes in the updating process, HTV can weaken the influence of error accumulation.

To investigate the effectiveness of the different verifiers, we implement a comparison between several variants of our tracker. Show the upper of Table III. We denote HTV without online updating as HTV-N. HTV-G is the version with general updating [6,12,16]. And HTV applying one of the two verifiers is called HTV-α and HTV-β, respectively.

As illustrated in the bottom of Table III, HTV-G obtains a slight improvement over HTV-N. In contrast, HTV-α and HTV-β take significant performance gains far greater than HTV-G does, which we attribute their effectiveness to the decrease of the error caused by inaccurate tracking predictions. Moreover, the combination of HTV-α and HTV-β can further enhance the performance, indicating the complementarity between α and β verifiers. In terms of tracking speed, our updating strategy takes a small cost, thereby enabling HTV a real-time running at 35 fps in FRHT.

Furthermore, we compare HTV with the latest work [24] of head tracking. The authors evaluate their method on five sequences of Headtracker with RMSE. We repeat the experiment and show the comparison in Table IV. HTV gains a smaller RMSE in all 5 sequences.

VI. CONCLUSION

In this paper, we propose the fully annotated benchmark FRHT for head tracking. FRHT provides the massive sequences of 360-degree head rotation as well as similar object clutter. These great advantages in real application make that FRHT can be taken as the standard benchmark to evaluate the performance of head trackers. Furthermore, we propose an online updating scheme with the combination of verifying strategies. By the prior color of observations and the template reappearance, the proposed method can effectively prevent the drift problem.

ACKNOWLEDGMENT

This work is supported in part by National Basic Research Program of China (973 Program): 2015CB351802, and Natural Science Foundation of China (NSFC): 61390515, 61390511, 61572465 and 61650202.

REFERENCES

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. Torr, "Fully-convolutional siamese networks for object tracking," in *ECCV*, 2016, pp. 850–865.

[2] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *CVPR*, 1998, pp. 232–237.

[3] N. Bouaynaya, W. Qu, and D. Schonfeld, "An online motion-based particle filter for head tracking applications," in *ICASSP*, 2005, pp. 225–228.

[4] Z. Chen, Z. Hong, and D. Tao, "An experimental survey on correlation filter-based tracking," *Computer Science*, vol. 53, no. 6025, pp. 68–83, 2015.

[5] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *CVPR*, 2017.

[6] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *ICCV*, 2015, pp. 4310–4318.

[7] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: learning continuous convolution operators for visual tracking," in *ECCV*, 2016, pp. 472–488.

[8] 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 *CVPR*, 2015.

[9] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *ICCV*, 2017.

[10] Y. Li, H. Ai, C. Huang, and S. Lao, "Robust head tracking with particles based on multiple cues fusion," in *ECCV*, 2006, pp. 29–39.

[11] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *CVPR*, 2017.

[12] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *ICCV*, 2015, pp. 3074–3082.

[13] K. Matej, M. Jiri, L. Ales, and et al., "A novel performance evaluation methodology for single-target trackers," *TPAMI*, vol. 38, no. 11, pp. 2137–2155, 9 2016.

[14] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *CVPR*, 2017.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS Workshop*, 2017.

[16] J. X. M. Z. W. H. Qiang Wang, Jin Gao, "Dcfnet: Discriminant correlation filters network for visual tracking," 2017.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[18] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *CVPR*, 2017.

[19] T. Vu, A. Osokin, and I. Laptev, "Context-aware CNNs for person head detection," in *ICCV*, 2015, pp. 2893–2901.

[20] Y. Wu, J. Lim, and M. Yang, "Object tracking benchmark," *TPAMI*, vol. 37, no. 9, pp. 1834–1848, 2015.

[21] K. Yachi, T. Wada, and T. Matsuyama, "Human head tracking using adaptive appearance models with a fixed-viewpoint pan-tilt-zoom camera," in *FG Workshops*, 2000, pp. 150–155.

[22] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Young Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *CVPR*, 2017.

[23] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: robust tracking via multiple experts using entropy minimization," in *ECCV*, 2014, pp. 188–203.

[24] X. Zhang, W. Hu, H. Bao, and S. Maybank, "Robust head tracking based on multiple cues fusion in the kernel-bayesian framework," *TCSVT*, vol. 23, no. 7, pp. 1197–1208, 2013.