Performance Evaluation of Visual Tracking Algorithms on Video Sequences With Quality Degradation

Fang, Yuming; Yuan, Yuan; Li, Leida; Wu, Jinjian; Lin, Weisi; Li, Zhiqiang

2017

Fang, Y., Yuan, Y., Li, L., Wu, J., Lin, W., & Li, Z. (2017). Performance Evaluation of Visual Tracking Algorithms on Video Sequences With Quality Degradation. IEEE Access, 5, 2430-2441.

https://hdl.handle.net/10356/86821

https://doi.org/10.1109/ACCESS.2017.2666218

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2017; accepted January 26, 2017, date of publication February 8, 2017, date of current version March 15, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2666218

Performance Evaluation of Visual Tracking Algorithms on Video Sequences With Quality Degradation

YUMING FANG[1], (Member, IEEE), YUAN YUAN[2], LEIDA LI[3], (Member, IEEE), JINJIAN WU[4], (Member, IEEE), WEISI LIN[2], (Fellow, IEEE), AND ZHIQIANG LI[1]

[1]School of Information Technology and School of Statistics, respectively, Jiangxi University of Finance and Economics, Nanchang 330032, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798
[3]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China
[4]School of Electronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Leida Li (reader1104@gmail.com)

This work was supported part by the National Natural Science Foundation of China under Grant 61571212 and Grant 71663024, in part by the Natural Science Foundation of Jiangxi under Grant 20161ACB21014, and in part by the Qinglan Project of Jiangsu Province.

ABSTRACT Recently, there are lots of visual tracking algorithms proposed to improve the performance of object tracking in video sequences with various real conditions, such as severe occlusion, complicated background, fast motion, and so on. In real visual tracking systems, there are various quality degradation occurring during video acquisition, transmission, and processing. However, most existing studies focus on improving the accuracy of visual tracking while ignoring the performance of tracking algorithms on video sequences with certain quality degradation. In this paper, we investigate the performance evaluation of existing visual tracking algorithms on video sequences with quality degradation. A quality-degraded video database for visual tracking (QDVD-VT), including the reference video sequences and their corresponding distorted versions, is constructed as the benchmarking for robustness analysis of visual tracking algorithms. Based on the constructed QDVD-VT, we propose a method for robustness measurement of visual tracking (RMVT) algorithms by accuracy rate and performance stability. The performance of ten existing visual tracking algorithms is evaluated by the proposed RMVT based on the built QDVD-VT. We provide the detailed analysis and discussion on the robustness analysis of different visual tracking algorithms on video sequences with quality degradation from different distortion types. To visualize the robustness of visual tracking algorithms well, we design a robustness pentagon to show the accuracy rate and performance stability of visual tracking algorithms. Our initial investigation shows that it is still challenging for effective object tracking for existing visual tracking algorithms on video sequences with quality degradation. There is much room for the performance improvement of existing tracking algorithms on video sequences with quality degradation in real applications.

INDEX TERMS Performance evaluation, quality degradation, robustness analysis, visual tracking, benchmarking.

I. INTRODUCTION

Visual tracking is a hot topic in the research areas of computer vision and multimedia processing. It can be widely used in various multimedia applications such as visual surveillance, robot navigation, medical image, human-computer interaction, *etc*. With the initial state of an object in the first frame of a video sequence, visual tracking algorithms aim to accurately predict the object states in the following video frames. Previously, there have been various visual tracking algorithms proposed for object tracking in video sequences with various conditions in tracking circumstances such as severe occlusion, complicated background, fast motion, *etc*. [27], [28].

In previous studies, there are also various video databases built as the benchmarks for performance evaluation of visual tracking algorithms [27], [28]. These databases are mainly composed of video sequences with various tracking challenges such as occlusion, complicated background, fast motion, *etc*. However, video quality degradation is rarely considered in these existing studies. In real multimedia systems,

there might be different distortion types involved in video sequences during video acquisition, compression, processing, etc. When video sequences are acquired with different light conditions, there might be contrast distortion and noises generated. With the limited bandwidth resources, video sequences have to be compressed during transmission and this would cause compression distortion. Thus, video sequences might be distorted due to different circumstances in real multimedia systems. For video sequences with quality degradation, the targets might be not tracked as accurately as those in good-quality video sequences. Therefore, the influence of video quality degradation on visual tracking should be investigated for the design of robust visual tracking algorithms.

In the past decades, the quality/performance evaluation methods has been widely investigated for various multimedia applications [1]–[4]. Early signal fidelity metrics such as SNR (singal-to-noise rate), PSNR (peak SNR), MAE (mean absolute error), and MSE (mean square error) are designed to estimate the image/video quality by simply comparing the distorted content with the reference one. These metrics can not obtain promising performance in visual quality assessment, since they do not take the visual content into account during quality prediction [1], [2]. To better predict the quality of visual signals, there are many perceptual metrics proposed recently, including SSIM (structure similarity) [5], VIF (visual information fidelity) [6], VSNR (visual signal-to-noise ratio) [7], IGM (internal generative mechanism) inspired metrics [8], [9], gradient similarity metric [13], etc.

Recently, with the requirement of many emerging multimedia applications, there are some studies focusing on proposing application-specific evaluation methods for specific visual content and visual processing algorithms. With the emerging interests in 3D visual content several years ago, there have been many studies investigating quality evaluation of 3D visual content [14]–[16]. Some studies also investigate the visual quality assessment for screen content images [12], [17], tone-mapped images [19], contrast-distorted images [11], [20], image blurring [10], image sharpness [26], *etc*. [4]. Besides the quality evaluation for specific types of images, there are also some studies investigating the performance evaluation of specific visual processing algorithms, including image retargeting [18], [25], image fusion [22], pedestrian detection [21], etc.

As introduced previously, there are currently some studies building video databases for performance evaluation of visual tracking algorithms [27], [28]. However, these studies mainly focus on investigating the robustness of visual tracking on different challenges from occlusion, complicated background, fast motion, *etc*. They do not take the video quality into account during the performance evaluation of visual tracking. Recently, there are several studies conducting experiments to evaluate the performance of face detection and event detection algorithms in video sequences with quality degradation [23], [24]. These studies show that the detection algorithms can obtain high detection accuracy in good-quality video. It has also demonstrated that the detection accuracy would decrease with video quality degradation [23]. For visual tracking algorithms, there is still no study systemically analyzing the influence of video quality on the performance of visual tracking algorithms. Thus, it is much desired to investigate the performance evaluation of visual tracking algorithms on video sequences with quality degradation.

In this study, we aim to carry out the first in-depth study on performance evaluation of visual tracking algorithms with quality degradation video. A Quality-Degraded Video Database for Visual Tracking (QDVD-VT), including 4 original video sequences and 40 distorted versions, is constructed as the benchmark for performance evaluation of visual tracking algorithms. Ten existing visual tracking algorithms published recently are chosen to conduct the experiments for robustness analysis. We define the metric called robustness measurement for visual tracking (RMVT) with video quality degradation to predict the robustness of different visual tracking algorithms for video sequences. In the proposed RMVT, both the visual tracking accuracy and stability on video sequences with quality degradation are considered. With the proposed RMVT, the performance of certain visual tracking algorithm with regarding tracking accuracy and stability can be obtained for different types of distortions. We also provide the in-depth analysis and discussion on how different distortions and their distortion levels influence the performance of visual tracking algorithms. To visualize the robustness of visual tracking algorithms well, we design a robustness pentagon to show the accuracy rate and performance stability of visual tracking algorithms. With the investigation in this study, we try to provide some possible research directions on visual tracking in the future. To the best knowledge of our knowledge, this is the first study to systematically investigate the performance evaluation of visual tracking algorithms on video sequences with quality degradation and QDVD-VT is the first related video database for visual tracking. Partial preliminary results of the study have been published in [35].

The reminder of this paper is organized as follows. In Section II, we introduce the process for the construction of the proposed QDVD-VT. Section III describes the details of the proposed method for robustness analysis of visual tracking algorithms. In Section IV, we conduct the experiments to evaluate the performance of different visual tracking algorithms. The final section concludes the work and provide our possible research in the future.

II. THE BENCHMARK

In this study, the QDVD-VT is built based on four reference video sequences from PROST [29], including the video sequences of *board*, *box*, *lemming*, and *liquor*. This database has been widely used in performance evaluation of existing visual tracking algorithms [27], [29]–[32]. In this database, the ground truth is labeled manually by the bound box of the target for visual tracking. The resolution of the video frames is 480 × 640. The video sequences are obtained by the fixed camera, which guarantees relatively stable quality

FIGURE 1. The video frame samples. The images in the first column are the reference video frames; the images in the second to the last column are the distorted versions. The distortion types from the first row to the last row are distortions from compression, contrast change, resolution, and white noise.

of video frames. In addition, the target size in each video sequences is constant. Thus, we can adjust the resolution of video sequences to investigate the influence of the varied target sizes on visual tracking performance. We provide some samples of the reference video frames in the first column of Fig. 1.

As indicated previously, video sequences have to be compressed due to the limited resource of storage. Additionally, we have to compress video sequences for efficient transmission. With video compression, the compression distortion would be brought into the video sequences. Thus, we use the compression distortion to generate the distorted video sequences in the proposed QDVD-VT. We give some distorted samples from compression distortion in the first row of Fig. 1.

During video acquisition, the environment variety would cause the luminance differences of video frames, which would bring in the distortion of contrast change. In the case where the video sequences are captured in bad conditions, such as rainy environment or night, the video sequences might suffer severe distortion of contrast change. In this study, we take the contrast change as one distortion type in the proposed QDVD-VT. We provide some distorted samples from contrast change in the second row of Fig. 1.

For various emerging devices, the video sequences have to be displayed in various display screens with different sizes. Furthermore, the camera sensors might also cause the video sequences with different sizes. The performance of visual tracking algorithms might be influenced on video sequences with different resolutions. Thus, we adjust the resolutions of video sequences as one factor in the proposed QDVD-VT. Some samples of video frames with different resolutions are shown in the third row of Fig. 1.

Noise is a common distortion type in video sequences. It might be brought in video sequences during video acquisition, processing, and other procedures. Here, we also consider the noise as one type of distortion in the proposed QDVD-VT. Some samples of noise-distorted video frames are provided in the fourth row of Fig. 1.

The other factor we take into account in the proposed QDVD-VT is the frame rate. Generally, visual tracking algorithms try to track the target in the current frame depending on the target features of the previous frames. With different frame rates, the dependency of previous frames might be different for the tracking accuracy of the current frame.

Totally, we take five different distortion types for video sequences in the proposed QDVD-VT: compression distortion, contrast change, resolution variety, white noise, and

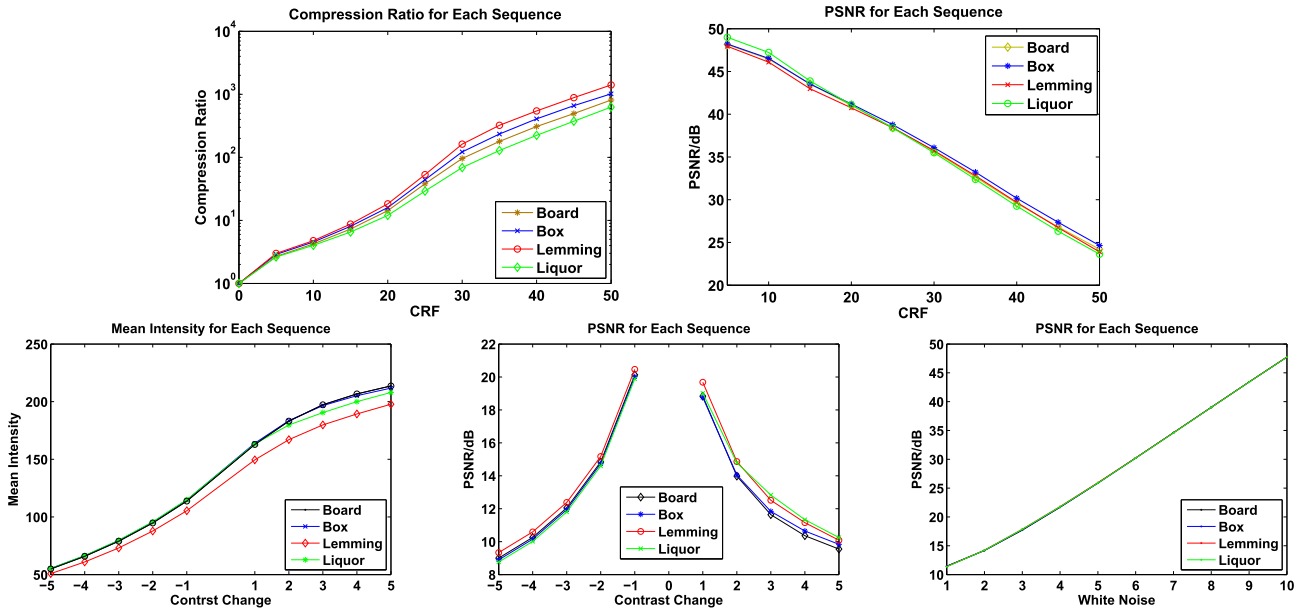

FIGURE 2. The properties of video sequences with different distortion types. In the fourth subfigure for the contrast change, the contrast parameter represented by x-axis values are $[1.2^{-5}, 1.2^{-4}, \ldots, 1.2^{5}]$. In the last subfigure for the white noise, the parameter σ represented by the x-axis values are $[0.6^{1}, 0.6^{2}, \ldots, 0.6^{10}]$.

frame rate of video sequences. With each type of distortions, we obtain different video quality levels to evaluate the robustness of visual tracking algorithms.

By using five distortion types, we create 48 distortion versions for each reference video sequence. Thus, there are $48 \times 4 + 4 = 196$ video sequences including four reference video sequences and 192 distorted versions totally in QDVD-VT. In the following, we analyze the statistics of the distorted video sequences in detail.

A. COMPRESSION DISTORTION

The compressed versions of video sequences are generated by using different values of constant rate factor (CRF) in H.264 codec. CRF is an important parameter in H.264 codec to encode video sequences with different bit rates. With increasing CRF values, the quality of video sequences would be degraded. In QDVD-VT, we generated the compression version of each video sequence by encoding it with 10 quality levels with the CRF in change of [5, 50]. In this study, we use ffmpeg [33] to encode the video sequences. We provide the compression ratio of each distorted video sequence in the first subfigure of Fig. 2. From this subfigure, we can see that the compression ratio changes from around 1 to around 10^3 between the reference video sequences and their distorted versions. Besides, the peak signal noise ratio (PSNR) of each distorted sequence is computed and shown in the second subfigure of Fig. 2. The PSNR values change from about 50 to 25 with different compression ratios.

B. CONTRAST CHANGE

Similarly, we generate the distorted versions for each video sequence with 10 levels of contrast change. For these 10 levels, there are five low and five high brightness levels for contrast change. The video sequences with contrast change are created by multiplying the video frames with a scaling factor s. For low contrast video sequences, the values of the scaling factor are smaller than 1, while for high contrast video sequences, the values of the scaling factor are larger than 1. The mean intensity value of each brightness level is given in the third subfigure in Fig. 2. Correspondingly, we also show the PSNR values for each video sequence in the fourth subfigure of Fig. 2. From this subfigure, we can see that the PSNR values decrease with the contrast changes to both lower and higher brightness levels.

C. RESOLUTION

The resolution variation can be generated to meet the low bandwidth limitation in H.264 codec. Here, we create 9 distorted versions for each video sequence with low resolutions by using the codec of ffmpeg setting different resolution parameters [33]. When coding video streams with ffmpeg, the resolution of video sequences can be adjusted by setting different resolution parameters. The resolution of video sequences is reduced from the original size (the reference video sequences) to the one ninth of the original size (the distorted versions with the lowest resolution). With lower resolutions of video sequences, the object sizes in video sequences would become smaller and this would influence the performance of visual tracking algorithms.

D. WHITE NOISE

In this study, the additive white noise is generated by a zero-mean Gaussian noise. There are 10 levels of Gaussian noise used to create the distorted versions of each

video sequence, where the Gaussian kernel σ varies in the range of $[0.6^1, 0.6^2, \ldots, 0.6^{10}]$. Correspondingly, the PSNR value changes from around 12 dB to around 48 dB, as shown in the fifth subfigure of Fig. 2. The PSNR values are highly dependent on σ and they are similar for different reference video sequences.

E. FRAME RATE

We also create the distorted versions of each video sequence with different frame rates. Totally, there are 9 levels for the varying frame rates of distorted video sequences. The frame rates vary from 30 FPS (frames per second) for the reference video sequences to 3 FPS for the distorted video sequences.

III. PROPOSED METHOD FOR ROBUSTNESS ANALYSIS

The accuracy rate of target tracking on video sequences can be used for performance evaluation of visual tracking algorithms. Besides, we also consider the stability of target tracking with video quality degradation for the robustness analysis for visual tracking algorithms. In this study, we propose the method called RMVT to evaluate the robustness of visual tracking algorithms from two aspects of accuracy rate and performance stability.

A. ACCURACY RATE EVALUATION

For accuracy rate evaluation, we use the bounding box overlap to measure the accuracy rates of visual tracking algorithms, which is widely used in performance evaluation of visual tracking algorithms [27], [31], [34]. Given the tracking bounding box B_t and the ground truth bounding box B_g, we calculate the overlap rate as follows.

$$R = \frac{Area(B_t \bigcap B_g)}{Area(B_t \bigcup B_g)} \tag{1}$$

where $\bigcap$ and $\bigcup$ denote the intersection and union of these two bound boxes; the function *Area* denotes the region area, which is represented by the number of pixels in the region.

B. PERFORMANCE STABILITY EVALUATION

The accuracy rate is undoubtedly an important measurement for performance evaluation of visual tracking algorithms. It is widely used for performance evaluation in existing visual tracking studies. However, in practical applications, the performance stability with video quality degradation is also an important factor for performance evaluation of visual tracking algorithms, since video quality might be degraded during acquisition, transmission, processing, *etc*. Besides, with similar accuracy rate for two different visual tracking algorithms, the performance stability provides another significant dimension for performance comparison. In this study, we propose a performance stability evaluation method through analyzing the performance of visual tracking algorithms.

With the five distortion types used in this study, the performance stability S is defined as a five-dimension vector as follows.

$$S = (S_{cd}, S_{cc}, S_{rd}, S_{wn}, S_{fr}), \tag{2}$$

where S_{cd}, S_{cc}, S_{rd}, S_{wn}, and S_{fr} represent the performance stability of visual tracking to compression distortion, contrast change, resolution distortion, white noise, and frame rate, respectively.

Generally, the tracking accuracy rate of visual tracking algorithms would change with the video quality degradation. In this study, we use the following two criteria to predict the accuracy change with video quality degradation to measure the performance stability of visual tracking algorithms: (1) *Rate of Accuracy Change*. For a robust visual tracking algorithm, the accuracy change should be slow with the decrease of video quality. (2) *Monotonicity*. For a robust visual tracking algorithm, the accuracy should show the monotonic change (degradation) with the decrease of video quality.

Given a reference video sequence and a list of its distorted versions from certain specific distortion type (such as compression distortion, white noise, etc.), we can calculate the accurate rates of any visual tracking algorithm on the reference video sequence A_r and its distorted versions $\{A_i : i = 1, 2, \ldots, N_k\}$, where N_k denotes the number of distorted video sequences from distortion type k. For the *ith* distorted video sequence, we can calculate the rate of accuracy change D_i as follows.

$$D_i = min\{1, \frac{|A_i - A_r|}{A_r}\} \tag{3}$$

where A_i and A_r represent the accuracy rates on the *ith* distorted and the reference video sequences. From Eq. (3), we can see that D_i is positive and it is correlated to the difference between A_i and A_r. If A_i is closer to A_r, D_i would be less, which demonstrates that the performance of the visual tracking algorithm would not change largely with the decrease with video quality. And thus, this algorithm can be regarded to be somewhat robust from *rate of accuracy change*.

In addition, the *monotonicity* of accuracy change for visual tracking algorithms can be computed as follows.

$$M_i = \begin{cases} 0 & A_i \leq A_{i-1} \\ min\{1, \frac{|A_i - A_{i-1}|}{A_{i-1}}\} & A_i > A_{i-}, \end{cases} \tag{4}$$

where A_i and A_{i-1} denote the tracking accuracy in the ith and $(i-1)$th distorted video sequences, respectively. Please note that we rank the distorted video with quality degrading for the ith and $(i-1)$th distorted video sequences, which means that the quality of the A_ith video sequence is worse than that of the A_{i-1} video sequence. For the distortion type of *Contrast Change*, we rank the distorted video sequences by two groups (high and low brightness) separately. From Eq. (4), we can see that the visual tracking algorithm with lower M_i would be more robust from the aspect of *monotonicity*.

According to the two criteria computed in Eqs. (3) and (4), we calculate the robustness of visual tracking algorithms for the distorted video sequence with distortion type k by fusing

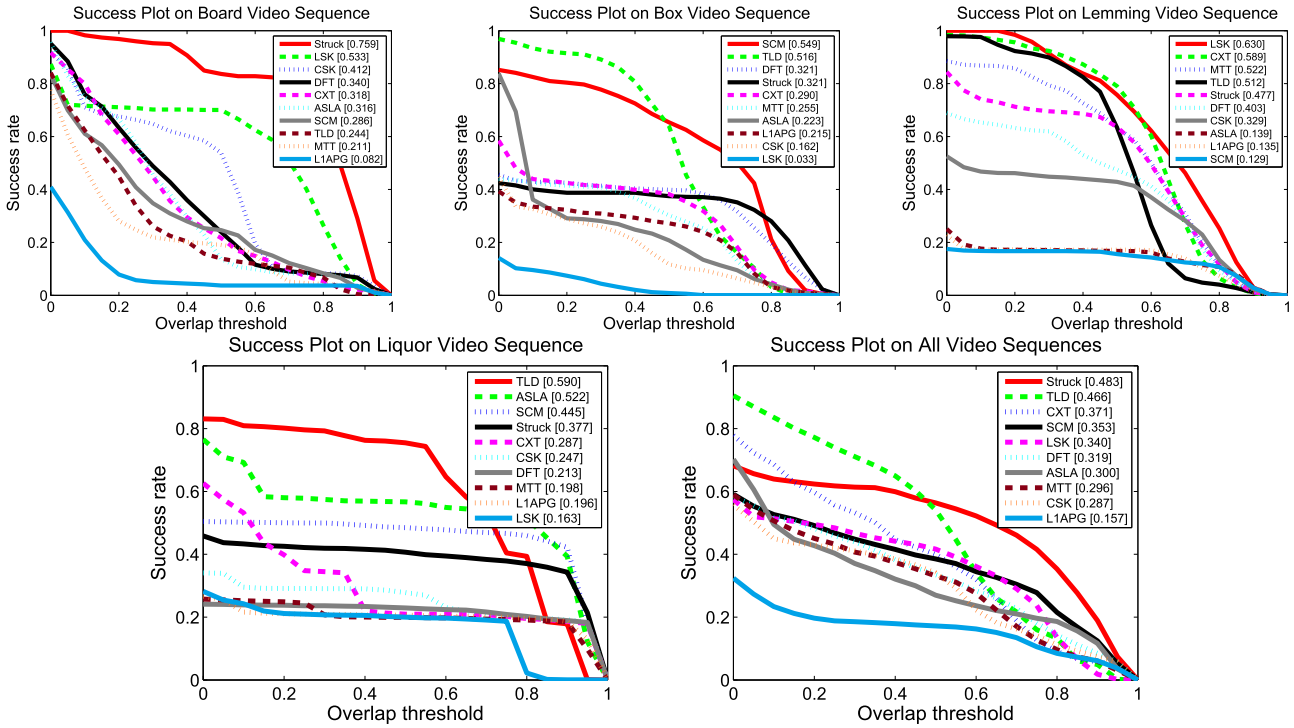

FIGURE 3. The experimental results of the different visual tracking algorithms on the reference video sequences. The first subfigure to the fifth subfigure: the performance of different visual tracking algorithms on *Board, box, lemming, liquor* video sequences and all the four reference video sequences.

these two criteria as follows.

$$S_k = \frac{1}{N_k}\sum_{i=1}^{N_k}(\alpha D_i + (1-\alpha)M_i) \tag{5}$$

where α is a parameter to weight the two components D_i and M_i; $k \in \{cd, cc, fr, rd, wn\}$ represents specific distortion type in Eq. (2). By considering both D_i and M_i in Eq. (5), we can see that the visual tracking algorithm will be more robust with lower value of S_k.

IV. EXPERIMENTAL RESULTS

In this section, we conduct the comparison experiment by using some existing visual tracking algorithms on QDVD-VT. The comparison results from these existing visual tracking algorithms are given. Besides, we also provide the in-depth analysis and discussion for the comparison experiment.

A. THE USED VISUAL TRACKING ALGORITHMS FOR ROBUSTNESS ANALYSIS

For the robustness evaluation of visual tracking algorithms on the constructed database, ten visual tracking algorithms are adopted to conduct the comparison experiment: ASAL [31], SCM [32], Struck [34], CXT [36], CSK [37], LSK [38], DFT [39], L1APG [40], MTT [41], TLD [42]. We select these visual tracking algorithms in this comparison experiment, since these algorithms can obtain better performance than other existing ones on a large-scale database, as shown in the study [27]. These visual tracking algorithms were published in recent years. We obtained the source code of these algorithms from the study [27].

Generally, there are the following key components in most visual tracking algorithms: target representation, search mechanism, and model update [27]. Here, we introduce the used visual tracking algorithms in these three aspects briefly. Some classical works use the template (raw intensity values) for target representation in visual tracking algorithms, such as DFT and CSK. Besides the template feature, visual tracking algorithms also use Haar-like feature (Struck), binary pattern (TLD), *etc*. Sparse representation is also widely used for object representation in existing visual tracking algorithms, including ASLA, SCM, L1APG, MTT, LSK, *etc*. Generally, there are two methods to predict the state of the target objects: the deterministic and stochastic methods. Within the optimization framework of visual tracking, the target can be located by local optimum search (LSK and DFT). To address the local minima problem, dense sampling methods are used in target searching (TLD, Struck, CSK, and CXT). Stochastic search methods such as particle filters are also widely used in visual tracking algorithms such as ASLA, SCM, L1APG, and MTT. It is important for visual tracking algorithms to update the target representation for appearance variations. Most of recent visual tracking algorithms used the model update process during visual tracking.

B. ACCURACY RATE ON REFERENCE VIDEO SEQUENCES

In this subsection, we provide the *accuracy rates* of the used visual tracking algorithms on the reference video sequences. The experimental results of different visual tracking algorithms are shown in Fig. 3. From this figure, we can see that the performance of different visual tracking algorithms varies

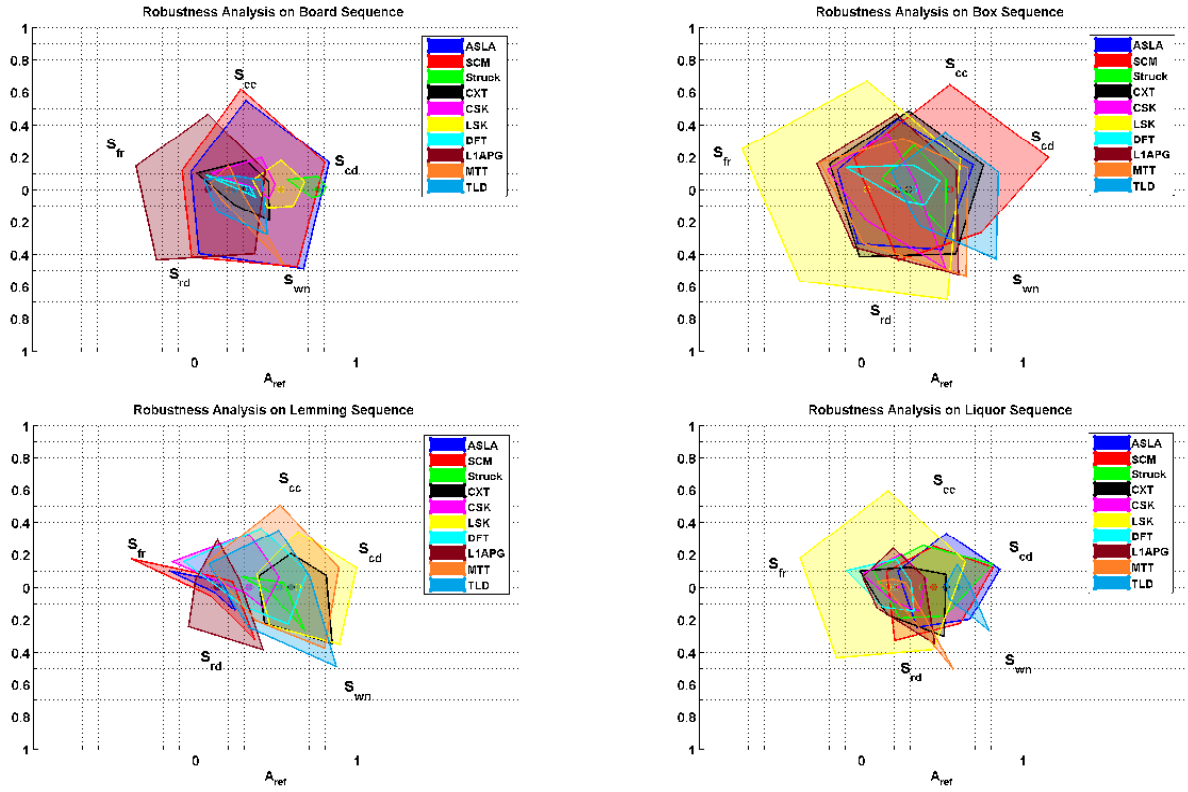

FIGURE 4. The robustness results of the different visual tracking algorithms on the video sequences. The first subfigure to the fourth subfigure: the robustness performance of different visual tracking algorithms on *Board, box, lemming,* and *liquor* video sequences.

on different video sequences. For *board* video sequence, Struck can obtain the best performance among the used visual tracking algorithms, while SCM, LSK, and TLD can obtain the best performance on the video sequences of *box*, *lemming* and *liquor*, respectively. From the first to the fourth subfigures, we can observe that the performance of these four trackers (Struck, SCM, LSK, and TLD) always rank ahead on different video sequences. The overall performance shown in the fifth subfigure in Fig. 3 demonstrates that Struck can obtain the best tracking performance among the used visual tracking algorithms. Also, SCM, LSK and TLD rank ahead for the tracking performance on all four reference video sequences. This demonstrates that the overall performance of visual tracking algorithms is stable on different video sequences.

C. OVERALL ROBUSTNESS ANALYSIS

As indicated in the previous section, the *robustness* of visual tracking algorithms can be evaluated by considering the *Accuracy Rate* (A_{ref}) on the reference video and *Stability* (S) with five distortion types. When comparing two visual tracking algorithms, the value of A_{ref} is firstly considered. If the accuracy rates of these two visual tracking algorithms are different largely, the visual tracking with higher accuracy rate is more preferable and the *Stability* is taken less consideration. In the case where the accuracy rates of two visual tracking algorithms are similar, the *Stability* S is important for robustness evaluation.

With the above analysis, we propose the *robustness pentagon* for robustness performance evaluation of visual tracking algorithms on video sequences, as shown in Fig. 4. In these figures, the robustness performance of each visual tracking algorithm is represented by a pentagon, whose center point denotes the *accuracy rate* of the visual tracking algorithm for the reference video. The *stability* from each distortion type is denoted by the distance from each vertex to center point of the pentagon, as demonstrated in Fig. 4. For the center of each pentagon, we use the value of x-axis to represent the *accuracy rate* of the corresponding visual tracking algorithm. Thus, the accuracy rate difference between two visual tracking algorithms can be represented by the distance between two robustness pentagon. With similar accuracy rate values from two visual tracking algorithms, the center points of these two pentagons are close and we can compare the S values denoted by the vertexes of these two overlapped pentagons.

From the first subfigure in Fig. 4, we can observe that the pentagon of Struck is in the right-most, which demonstrates that the accuracy rate of Struck for the reference *Board* video sequence is highest among the compared visual tracking algorithms. The small pentagon also shows that *Struck* is one of the most stable algorithms among the compared visual tracing algorithms. LSK can also obtain good performance on *Board* video sequence and it is more stable compared with CSK and ALSA. From the second subfigure of Fig. 4, we can see that the pentagons of SCM and TLD are right-most, which

shows that the accuracy rates of these two algorithms on *Box* video sequence are highest among the compared algorithms. However, the small pentagons of DFT and Struck demonstrate that these two algorithms are more stable than other tracking algorithms, while the large pentagon of LSK which is left-most demonstrates that LSK is the worst in robustness of tracking performance among the compared visual tracking algorithms. From the third subfigure of Fig. 4, we can observe that the pentagon of LSK is right-most, and thus, the accuracy rate of LSK is highest among the compared algorithms. Struck can also obtain a relative high accuracy rate with stable tracking performance (the pentagon of Struck is small). From the fourth subfigure of Fig. 4, we can observe that TLD is most robust among the compared visual tracking algorithms, since the pentagon of TLD is in the right-most and smallest among the compared visual tracking algorithms. In contrast, the left-most and small pentagon of LSK demonstrate the worst robustness in visual tracking among the compared visual tracking algorithms.

D. STABILITY WITH VARIOUS DISTORTIONS

1) STABILITY WITH COMPRESSION DISTORTION

We provide the accuracy rate vs. CRF curves of the visual tracking algorithms on the four video sequences in the first column of Fig. 5. From these experimental results, we can see that the Accuracy-CRF curves of most used visual tracking algorithms are not monotonic strictly. The accuracy rates of most tracking algorithms varies with different CRF values. For different visual tracking algorithms, the accuracy rates vary differently for various video sequences. From the first column of Fig. 5, we can observer that, for most visual tracking algorithms, the accuracy rates reduces with larger CRF values, which demonstrates that the accuracy rates of these tracking algorithms decrease with larger compression ratios. For *Board* video sequence, the accuracy rates of ASLA and SCM vary greatly with different CRF values, which means that ASLA and SCM are not stable with compression distortion on this video sequence. In contrast, the accuracy rates of other visual tracking algorithms on *Board* video sequence vary more slowly than ASLA and SCM, which means that these algorithms are more stable than ASLA and SCM on *Board* video sequence. The stability of all visual tracking algorithms on other video sequences can be also analyzed by the first column of Fig. 5. From the overall performance on all video sequences in the final subfigure of the first column in Fig. 5, we can see that the average accuracy rates of most visual tracking algorithms reduce with larger CRF values. This demonstrates that the accuracy rates of visual tracking algorithms reduce with increasing compression distortion.

2) STABILITY WITH CONTRAST CHANGE

The accuracy rate vs. contrast change curves of the used visual tracking algorithms on all four video sequences are shown in the second column of Fig. 5. From these curves, we can see that the Accuracy-Contrast Change curves of most existing visual tracking algorithms are not monotonic strictly. The accuracy rates of most visual tracking algorithms vary greatly with different contrast change. Specifically, the performance of ASLA and SCM changes greatly on video sequences *Board*, *Box*, and *Liquor*. The accurate rates of ASLA and SCM are low on the video sequence *Lemming*. The accurate rates of TLD and CXT vary greatly on video sequences *Box* and *Lemming*. From the last figure of the second column in Fig. 5, the overall performance of most visual tracking algorithms have no much change on video sequences with the distortion from different contrast change. This demonstrates that there is no much influence on performance of most visual tracking algorithms for video sequences with quality degradation from the contrast change within certain scope. This fruther demonstrates that most visual tracking algorithms are somewhat stable on video sequences with quality degradation from contrast change.

3) STABILITY WITH FRAME RATE

The accuracy rate vs. frame rate curves of the used visual tracking algorithms on all four video sequences can be found in the third column of Fig. 5. From the figures of this column, the performance of most visual tracking algorithms decreases with the smaller frame rates of the video sequences. There are some visual tracking algorithms whose performance varies largely with different frame rates for the video sequences. From the second figure of the third column in Fig. 5, the performance of L1APG and MTT changes greatly on video sequence *Box* with different frame rates. The performance of Struck and CXT are not stable on the video sequence *Liquor* with different frame rates. From the last figure of the third column in Fig. 5, we can observe that overall performance of most visual tracking algorithms decreases with quality degradation from smaller frame rates on all four video sequences.

4) STABILITY WITH RESOLUTION

We provide the accuracy rate vs. resolution curves of the used visual tracking algorithms in the first fives figures of Fig. 6. From the first, second and fourth figures in Fig. 6, we can observe that the accuracy rates of ALSA and SCM vary greatly on video sequences *Board*, *Box*, and *Liquor* with quality degradation from different resolutions, which demonstrate that the performance of ALSA and SCM is not stable on these video sequences with different resolutions. From the third figure of Fig. 6, the performance of LSK, TLD, and MTT is not stable on video sequence *Lemming* with quality degradation from different resolutions. From the fifth figure in Fig. 6, we can see that there is not much change for the overall performance of most visual tracking algorithms on all video sequences with quality degradation from different resolutions. This demonstrates that most visual tracking algorithms are somewhat stable on video sequences with quality degradation from different resolutions.

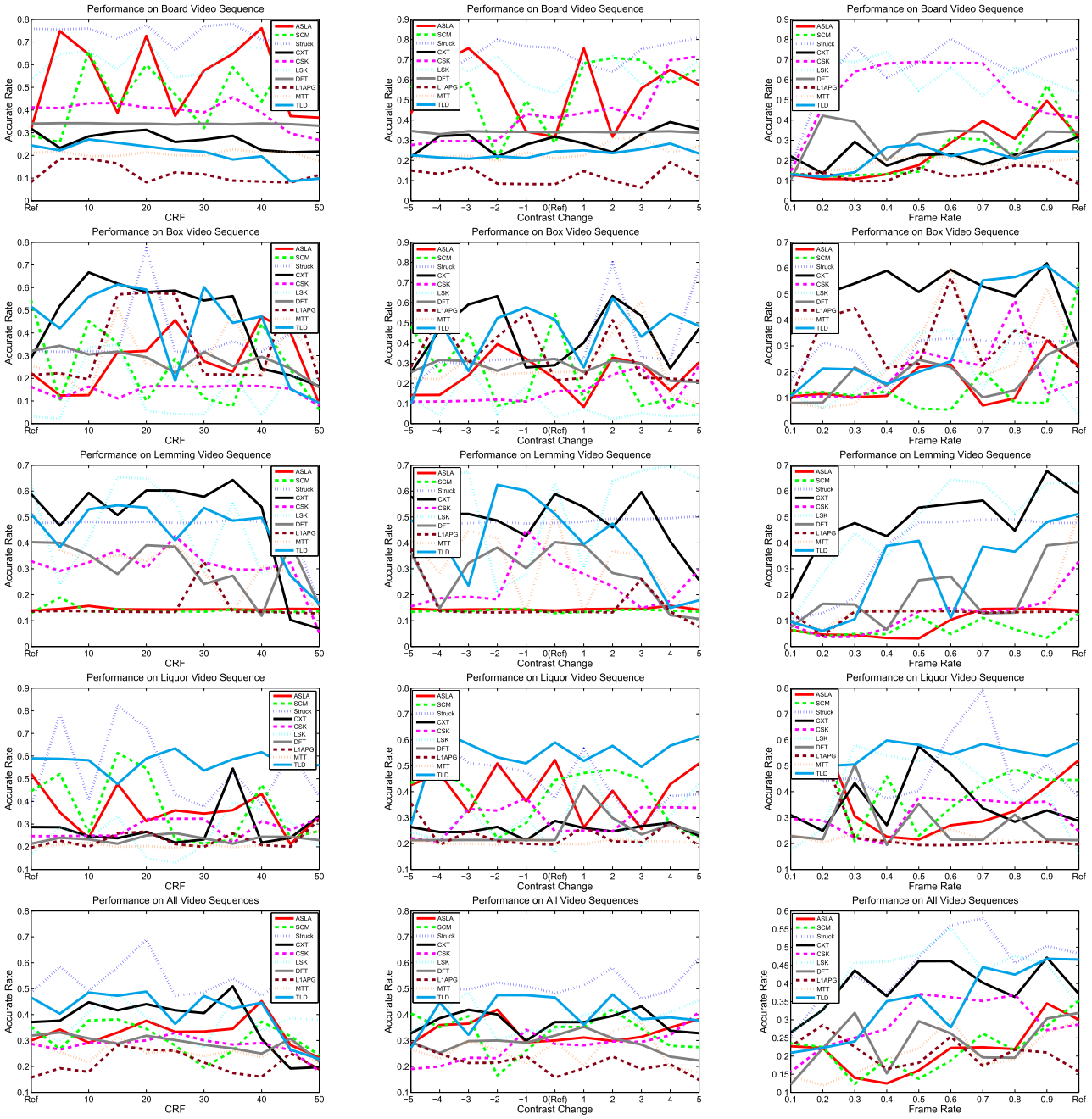

FIGURE 5. The experimental results of the different visual tracking algorithms on video sequences with quality degradation. The first column to the third column: experimental results on video sequences from distortions of *cd*, *cc* and *fr*.

5) STABILITY WITH WHITE NOISE

The accuracy rate vs. white noise curves of the used visual tracking algorithms are shown from the sixth to tenth figures in Fig. 6. From the sixth and seventh figures, we can see that the accuracy rates of ASLA change greatly for video sequences *Board* and *Box* with quality degradation from white noise, which demonstrates that the performance of ASLA is not stable for this video sequence with quality degradation from white noise. From the eighth figure of Fig. 6, we can see that the accuracy rates of TLD and LSK vary greatly on video sequence *Lemming* with quality degradation from white noise. This demonstrates that the performance of TLD and LSK is not stable on this video sequence with quality degradation from white noise. From the ninth figure in Fig. 6, we can observe that the performance of most visual tracking decreases on video sequence *Liquor* with larger distortion from white noise. From the last figure in Fig. 6, we can see that the overall accuracy rates of most visual tracking algorithms decreases on all video sequence with larger distortion from white noise, which demonstrates that the performance of most visual tracking algorithms are not much stable on video sequences with quality degradation from white noise.

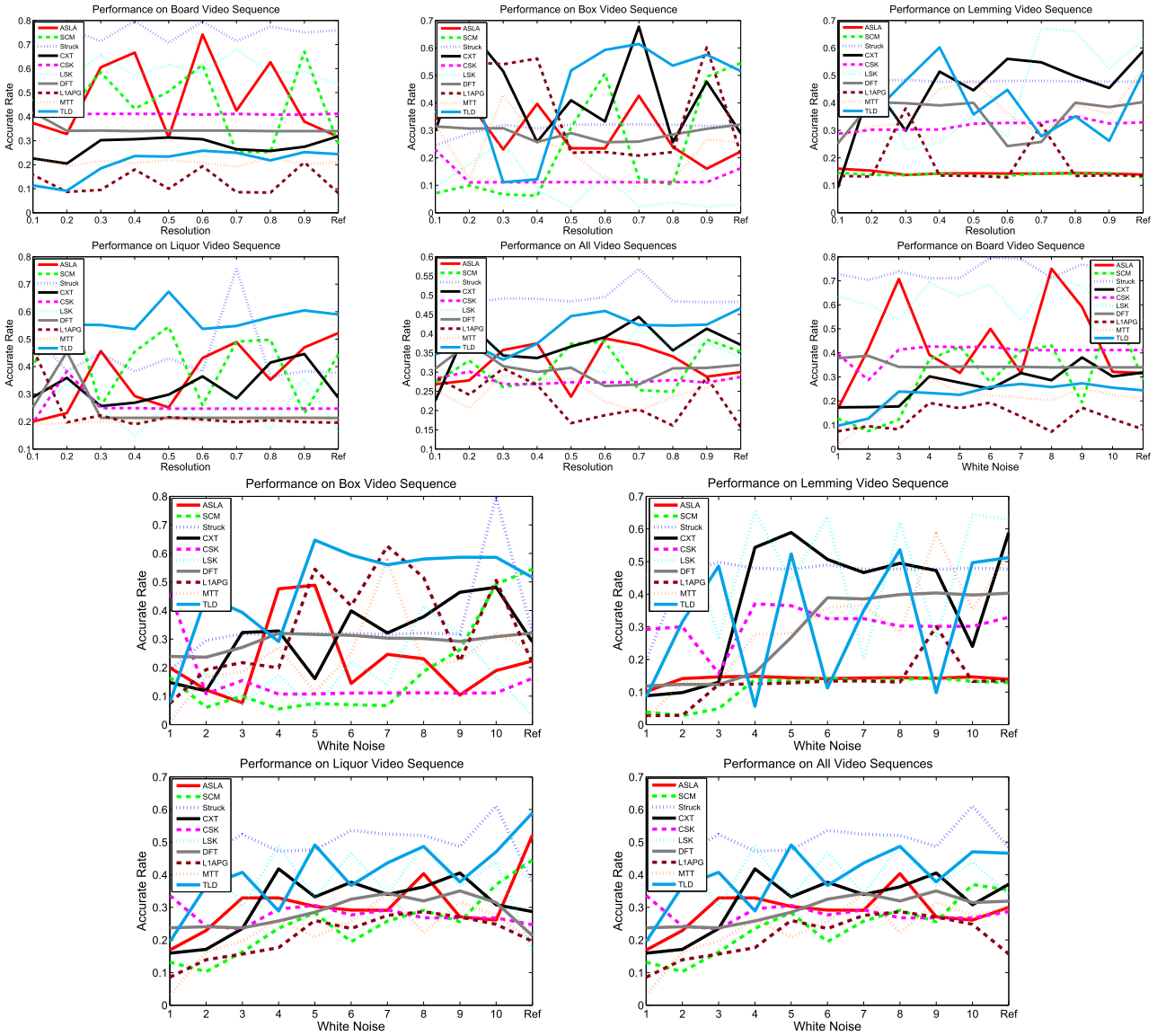

FIGURE 6. **The experimental results of the different visual tracking algorithms on video sequences with quality degradation. The first four subfigures: experimental results on video sequences with the distortion of *res*; the last four subfigures: experimental results on video sequences with the distortion of *wn*.**

V. DISCUSSION AND CONCLUSION

In this paper, we have investigated the robustness of visual tracking algorithms on video sequences with quality degradation. A new database of video sequences is constructed for the robustness analysis of visual tracking algorithms. In this database, five common distortion types from compression distortion, contrast change, resolution change, white noise and frame rate change are used to generate the distorted video sequences. A method called RMVT is proposed for robustness analysis of visual tracking algorithms based on the accuracy rate and performance stability. Besides, a pentagon representation is designed to visualize the robustness of visual tracking algorithms. In the experiments, we analyze the robustness of existing visual tracking algorithms on the video sequences with quality degradation from different distortion types in detail. From the experimental results in Fig. 4 5 6, we can observe that the accuracy rates of many existing visual tracking algorithms vary greatly on specific video sequences with quality degradation. Strictly speaking, there is no any visual tracking algorithm which is robust to all video sequences with quality degradation from different distortion types. Thus, the robust visual tracking algorithm is still much desired in the research community.

REFERENCES

[1] W. Lin and C. C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[2] Z. Wang and A. C. Bovik, "Modern image quality assessment," in *Syntheses Lectures on Image, Video and Multimedia Processing*. San Mateo, CA, USA: Morgan Kaufmann, Mar. 2006.

[3] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," in *Proc. ISRN Signal Process.*, 2013, pp. 1–53.
[4] Y. Fang, "Application-specific visual quality assessment: Current status and future trends," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2015, Art. no. 87.
[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[6] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[7] D, M. Chandler, and S. S. Hemami, "VSNR: A wavelet-based visual signalto-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[8] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43–54, Jan. 2013.
[9] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.
[10] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 39–50, Jan. 2016.
[11] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.
[12] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408–4421, Nov. 2015.
[13] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.
[14] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality prediction of asymmetrically distorted stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3400–3414, Nov. 2015.
[15] K. Moorthy and A. C. Bovik, "A survey on 3D quality of experience and 3D quality assessment," *Proc. SPIE*, vol. 8651, p. 86510M, Mar. 2013.
[16] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1940–1953, May 2013.
[17] H. Yang, Y. Fang, W. Lin, and Z. Wang, "Subjective quality assessment of screen content images," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2014, pp. 257–262.
[18] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "Comparative study of image retargeting," *ACM SIGGRAPH Asia*, vol. 29, no. 6, pp. 1–9, 2010.
[19] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.
[20] K. Gu, G. Zhai, X. Yang, W. Zhang, and M. Liu, "Subjective and objective quality assessment for images with contrast change," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 383–387.
[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[22] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[23] E. Kafetzakis, C. Xilouris, M. A. Kourtis, M. Nieto, I. Jargalsaikhan, and S. Little, "The impact of video transcoding parameters on event detection for surveillance systems," in *Proc. IEEE Int. Symp. Multimedia*, Sep. 2013, pp. 333–338.
[24] P. Korshunov and W. T. Ooi, "Video quality for face detection, recognition, and tracking," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7, no. 3, pp. 1–14, 2011.
[25] Y. Fang, K. Zeng, Z. Wang, W. Lin, Z. Fang, and C.-W. Lin, "Objective quality assessment for image retargeting based on structural similarity," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 4, no. 1, pp. 95–105, Jan. 2014.
[26] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, Jun. 2016.
[27] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2013, pp. 2411–2418.
[28] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.
[29] H. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, "PROST: Parallel robust online simple tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 723–730.
[30] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and K-selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1313–1320.
[31] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1822–1829.
[32] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2012, pp. 1838–1845.
[33] F. Bellard, M. Niedermayer. (2014). *Ffmpeg*. [Online]. Available: http://www.ffmpeg.org
[34] S. Hare *et al.*, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2011, pp. 2096–2109.
[35] Y. Fang, Y. Yuan, L. Xu, and W. Lin, "A benchmark for robustness analysis of visual tracking algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2016, pp. 1120–1124.
[36] T. B. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2011, pp. 1177–1184.
[37] F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2012, pp. 702–715.
[38] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust Tracking using Local Sparse Appearance Model and K-Selection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2011, pp. 2968–2981.
[39] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2012, pp. 1910–1917.
[40] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2012, pp. 1830–1837.
[41] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2012, pp. 2042–2049.
[42] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2010, pp. 49–56.

YUMING FANG received the Ph.D. degree from Nanyang Technological University, Singapore, the M.S. degree from the Beijing University of Technology, China, and the B.E. degree from Sichuan University in China. He is currently a Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, and 3D image/video processing. He serves as an Associate Editor of the *IEEE Access* and is on the Editorial Board of the *Signal Processing: Image Communication*.

YUAN YUAN received the B.E. degree in electronic engineering from the Beijing University of Post and Telecommunication, Beijing, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore. His current research interests include visual surveillance, object tracking, and visual attention.

LEIDA LI received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Visiting Ph.D. Student with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the School of Information and Control Engineering, China University of Mining and Technology, China, and a Senior Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include multimedia quality assessment, information hiding, and image forensics.

JINJIAN WU received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. In 2013, he was a Research Assistant with Nanyang Technological University, Singapore. From 2013 to 2014, he was a Post-Doctoral Research Fellow with Nanyang Technological University. From 2013 to 2015, he was a Lecturer with Xidian University. Since 2015, he has been an Associate Professor with the School of Electronic Engineering, Xidian University. His research interests include visual perceptual modeling, saliency estimation, quality evaluation, and just noticeable difference estimation. He has served as the TPC member of the ICME2014, the ICME2015, the PCM2015, and the ICIP2015. He was a recipient of the Best Student Paper of the ISCAS 2013.

WEISI LIN (F'15) received the Ph.D. degree from the King's College London, London University, U.K. He served as the Laboratory Head of Visual Processing, Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering. His technical expertise includes perceptual modeling and evaluation of multimedia signals, image processing and video compression, in which he has authored 160 journal papers and 230 conference papers, filed seven patents, authored two books, edited three books, and written nine book chapters. He is a Chartered Engineer of the IEEE, a fellow of the IET, and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He served as a Guest Editor of seven special issues in different scholarly journals. He has been a Technical Program Chair of the IEEE ICME 2013, the PCM 2012, the QoMEX 2014, and the VCIP 2017. He chaired the IEEE MMTC Special Interest Group on QoE 2012–2014. He has been a keynote/invited/panelist/tutorial speaker in over 20 international conferences and a Distinguished Lecturer the Asia-Pacific Signal and Information Processing Association, 2012–2013, the IEEE Circuits and Systems Society 2016–2017. He is currently an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE Transactions on CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the *Journal of Visual Communication and Image Representation*, and a past Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE SIGNAL PROCESSING LETTERS.

ZHIQIANG LI received the Ph.D. degree from the Jiangxi University of Finance and Economics. He is currently a Professor with the Jiangxi University of Finance and Economics. He has authored over 30 academic papers. His research interest includes intelligent optimization and machine learning.