A BENCHMARK FOR ROBUSTNESS ANALYSIS OF VISUAL TRACKING ALGORITHMS

Yuming Fang[1], *Yuan Yuan*[2], *Long Xu*[3], *and Weisi Lin*[2]

[1]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[3]Key Laboratory of Solar Activity, National Astronomical Observatories, CAS, China

ABSTRACT

In this study, we investigate the robustness of existing visual tracking algorithms with quality-degraded video. A video database including the reference video sequences and their distorted versions is created as the benchmark for robustness analysis of visual tracking algorithms. Ten existing visual tracking algorithms are used to conduct the experiments for robustness analysis based on the benchmark. Our initial investigation demonstrates that all the existing visual tracking algorithms cannot obtain the robust visual tracking results for quality-degraded video sequences. The experimental results in this study show that there is still much room for the design of robust visual tracking algorithms.

Index Terms— Robustness analysis, visual tracking, quality-degraded video, quality assessment.

1. 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. Given the initial state of a target in the first frame of one video sequence, visual tracking aims to accurately estimate the target states in the following frames of the video sequence. In the past decades, there have been various visual tracking algorithms proposed for object tracking in video sequences with a wide variety of conditions in tracking circumstances such as severe occlusion, complicated background, fast motion, etc. [11, 12].

Previously, there are also various video databases built as the benchmarks for performance evaluation of visual tracking algorithms [11, 12]. These databases mainly include video sequences with various tracking challenges such as occlusion, complicated background, fast motion, etc. However, the quality degradation in video sequences is rarely considered in these existing studies. In real systems, there might be different types of distortions involved in the video sequences during video acquisition, compression, processing, etc. For example, the contrast distortion and noises might be generated when video sequences are captured with different light. Due to the limited bandwidth resources, video sequences have to be compressed during transmission, which might cause compression distortion. Thus, the visual quality of video sequences would be degraded due to different circumstances in real multimedia systems. With quality degraded video sequences, the targets might be not tracked as accurately as those in good-quality video sequences. Therefore, the influence of quality-degraded video 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 studied for various multimedia applications [1, 2, 3, 4]. Early signal fidelity metrics such as SNR (singal-to-noise rate), PSNR (peak SNR), MAE (mean absolute error), MSE (mean square error), etc. are designed to predict the signal quality by 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 inforamtion fidelity) [6], VSNR (visual signal-to-noise ratio) [7], IGM (internal generative mechanism) inspired metrics [8, 9], gradient similarity metric [10], etc. However, these metrics are designed for visual quality assessment of visual content. They cannot be used for robustness analysis of visual tracking algorithms.

In this study, we aim to carry out the initial in-depth study on robustness analysis of visual tracking algorithms with quality-degraded video. A video database, including 4 original video sequences and 192 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. In this study, both the visual

This work was partially supported by NSFC (No.61571212,61202242) and NSF of Jiangxi Province (No.20151BDH80003).

Fig. 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.

tracking accuracy and stability on video sequences with different levels of distortions are considered. 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. With the initial investigation in this study, we also try to provide some possible research directions on visual tracking in the future. To the best of our knowledge, this is the first study to systematically investigate the performance evaluation of visual tracking algorithms with quality-degraded video and the constructed video database is the first related video database for robustness analysis of visual tracking.

2. THE BENCHMARK

The database is built based on four reference video sequences from PROST [13], including the video sequences of *board*, *box*, *lemming*, and *liquor*. These four video sequences are widely used in performance evaluation of visual tracking algorithms [11] [13]-[16]. The bound boxes of targets are manually labeled as the ground truth for visual tracking. The original resolution of these video sequences is 480×640. These video sequences are captured by the fixed camera, which guarantees relatively stable quality of the 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. Some samples of the reference video frames are given in the first column of Fig. 1.

With the limited resource of storage, video sequences should be compressed after acquisition. Additionally, video sequences have to be compressed for more efficiently transmission. However, it would bring into compression distortion. Thus, we include the compression distortion in the constructed database. During video acquisition, the environment variety might cause the luminance differences of video sequences, which would bring to the distortion of contrast change. For example, when the video sequences are captured in the night, rainy environment or other conditions, the video sequences might suffer severe distortion of contrast change. In this study, we take the contrast change as one distortion in the constructed database. With various emerging devices, the video sequences have to be displayed in display screens with different sizes. In addition, the camera sensors might also cause the video sequences with different sizes. With different resolutions, the performance of visual tracking algorithms might be influenced on video sequences. Thus, we adjust the resolutions of video sequences as one factor in the constructed database.

The noise is one type of common distortions in video sequences. It might be caused during video acquisition, processing, and other procedures. In this study, we consider the noise as one type of distortion in the constructed database. The other factor we take into account in this study 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 factors for video sequences in the proposed database: compression distortion, contrast change, resolution variety, white noise, and frame rate of video sequences. With each type of distortions, we obtain different video quality levels to evaluate the performance 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 the database. In the following, we will introduce the distorted video sequences in detail.

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 this study, we generated the compression version of each video sequence by encoding it with 10 quality levels with the CRF in change of

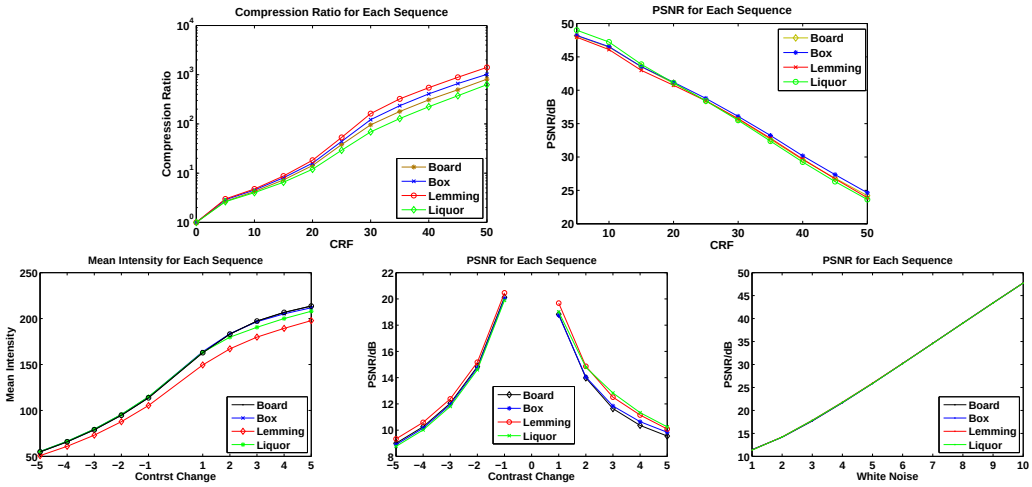

Fig. 2. The properties of video sequences with different distortion types. In the fourth subfigure for the contrast change, the contrast parameter represented by x-axis values are $[1.2^{-5}, 1.2^{-4}, ..., 1.2^{5}]$. In the last subfigure for the white noise, the parameter σ represented by the x-axis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

.

[5, 50]. Here, we use ffmpeg [17] to encode the video sequences. We provide the compression ratio of each distorted video sequence in the first subfigure of Fig. 2. Besides, the peak signal noise ratio (PSNR) of each distorted sequence is computed and shown in the second subfigure of Fig. 2.

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 mean intensity and PSNR values of each brightness level are given in the third and fourth subfigures of Fig. 2, respectively.

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 sequences with low resolutions by using the codec of ffmpeg [17]. The resolution is reduced from the original size(the reference video sequences) to one tenth of the original size (the distorted versions).

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 $[0.6^{1}, 0.6^{2}, ..., 0.6^{10}]$. Correspondingly, the PSNR value changes from around 12 dB to around 48 dB, as shown in the last subfigure in Fig. 2. The PSNR values are highly dependent on σ and they are similar for different reference video sequences.

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.

3. ROBUSTNESS ANALYSIS

The accurate rate of target tracking on video sequences can be used for performance evaluation of visual tracking algorithms. Besides the accurate rate of target tracking, we also consider the stability of target tracking with video quality degradation for the robustness analysis for visual tracking algorithms. Here, we provide the experimental results for the robustness analysis of visual tracking algorithms from two aspects of accuracy rate and performance stability.

3.1. Accurate Rate Evaluation

We use the bounding box overlap to measure the accurate rates of visual tracking algorithms, which is widely used in performance evaluation of visual tracking algorithms [11, 15, 18]. Given the tracking bounding box B_t and the ground truth bounding box B_g, we calculate the overlap rate as follows.

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

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.

3.2. Robustness Analysis

Generally, the tracking accurate rate of visual tracking algorithms would decrease with the quality degradation in video

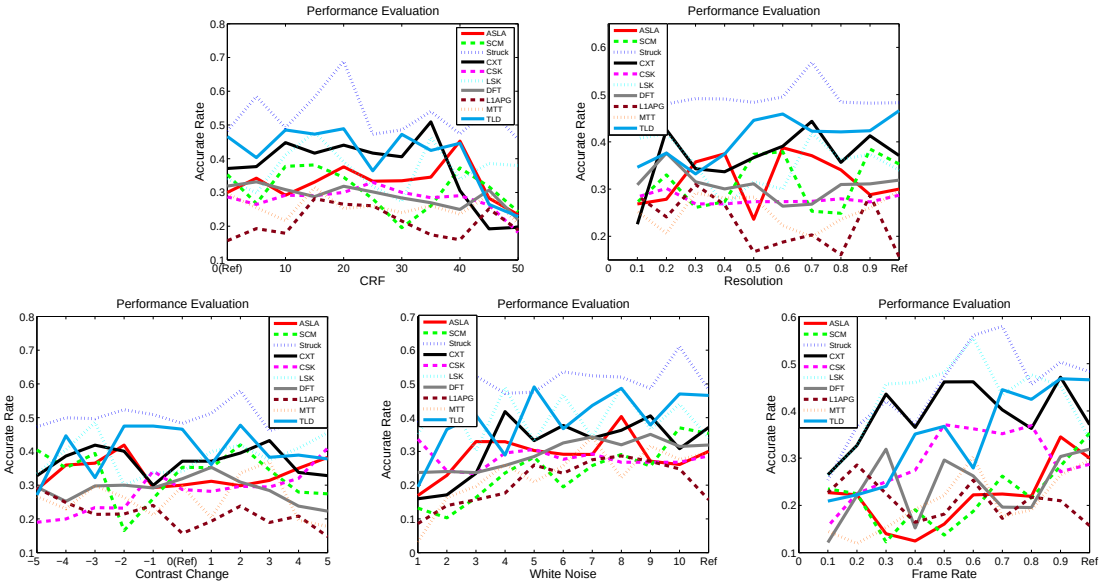

Fig. 3. The experimental results from different distortion types. 'Ref' denotes the performance of the reference video sequence. In the third subfigure for the contrast change, the contrast parameter represented by x-axis values are $[1.2^{-5}, 1.2^{-4}, ..., 1.2^{5}]$. In the fourth subfigure for the white noise, the parameter σ represented by the x-axis values are $[0.6^{1}, 0.6^{2}, ..., 0.6^{10}]$

.

sequences. In this study, we use the degradation rate of accuracy to analyze the robustness of visual tracking algorithms. 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 $\{A_i : i = 1, 2, ..., N_k\}$, where N_k denotes the number of distorted video sequences from distortion type k. To evaluate the robustness of visual tracking algorithms on the built database, we use ten trackers to conduct the comparison experiments: ASAL [15], SCM [16], Struck [18], CXT [19], CSK [20], LSK [21], DFT [22], L1APGT [23], MTT [24], TLD [25]. We use these trackers due to their better performance compared with other existing algorithms, as shown in [11].

The overall comparison experiment results based on the database are given in Fig. 3. We obtain the surprising results from the experiments. From the first subfigure, we can see that the performance would not decrease obviously with video quality degradation by compression distortion. When the compression distortion becomes larger (CRF varies from 30 to 40), surprisingly, the performance of all the trackers increase. From the second subfigure in Fig. 3, the performance of most trackers do not increase or decrease monotonously with the decreasing resolution. The similar experimental results are obtained when distortions of contrast change, white noise and frame rate exist in video sequences, as showed in Fig. 3.

It is really surprising that the tracking performance does not decrease with the video quality degradation. The reason might be that the structure information of visual content changes when there are distortions. As we know, most of the visual tracking algorithms are sensitive to the visual content changes. Therefore, it is unpredicted for the visual tracking algorithms whether they could work when there is some change of visual content. In addition, the performance of the existing tracking algorithms are not good enough, as shown in Fig. 3. We can see that the accurate rates of most algorithms are below 0.5, which might be another reason for the instability of existing visual tracking algorithms.

4. CONCLUSION

In this study, we provide the initial investigation for the robustness analysis of visual tracking algorithms. We created the first database of video sequences for visual tracking, including both reference and distorted video sequences. The experimental results show that most of the existing tracking algorithms cannot obtain robust performance for video sequences with quality degradation. The initial results demonstrate that there is still much room to design robust visual tracking algorithms. In the future, we will analyze the experiment results in-depth and try to provide more insightful results in the robustness analysis of visual tracking algorithms.

5. REFERENCES

[1] W. Lin, and C.C.J. Kuo. Perceptual visual quality metrics: a survey. *Journal of Visual Communication and Image Representation*, 22(4), 297-312, 2011.

[2] Z. Wang and A. C. Bovik. *Modern image quality assessment*. in syntheses lectures on Image, Video and Multimedia Processing, Morgan & Claypool Publishers, Mar. 2006.

[3] D. M. Chandler, Seven Challenges in Image Quality Assessment: Past, Present, and Future Research, *ISRN Signal Processing*, vol. 2013, Article ID 905685, 53 pages, 2013.

[4] Y. Fang, Application-specific visual quality assessment: current status and future trends. *International Conference on Internet Multimedia Computing and Service*, 2015.

[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(4), 600-612, 2004.

[6] H. R. Sheikh, and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, vol. 15 (2), 430-444, 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(9), pp. 2284-2298, 2007.

[8] J. Wu, W. Lin, G. Shi, and A. Liu. Perceptual Quality Metric With Internal Generative Mechanism. *IEEE Transactions on Image Processing*, 22(1): 43-54, 2013.

[9] K. Gu, G. Zhai, X. Yang, and W. Zhang, Using Free Energy Principle For Blind Image Quality Assessment. IEEE Transactions on Multimedia 17(1): 50-63, 2015.

[10] A. Liu, W. Lin, and M. Narwaria. Image Quality Assessment Based on Gradient Similarity. IEEE Transactions on Image Processing, 21(4): 1500-1512, 2012.

[11] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: a benchmark *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[12] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: an experimental survey. *IEEE Transactions on Pattern Analysis and Machine Integillence*, Vol. 36, No. 7, pp. 1442-1468, 2014.

[13] H. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, PROST: Parallel Robust Online Simple Tracking, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[14] B. Liu, J. Huang, L. Yang, and C. Kulikowsk. Robust Tracking using Local Sparse Appearance Model and K-Selection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[15] X. Jia, H. Lu, and M.-H. Yang. Visual Tracking via Adaptive Structural Local Sparse Appearance Model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] W. Zhong, H. Lu, and M.-H. Yang. Robust Object Tracking via Sparsity-based Collaborative Model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] F. Bellard, M. Niedermayer et al., Ffmpeg, http://www.ffmpeg.org, 2014.

[18] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured Output Tracking with Kernels. In ICCV, 2011.

[19] T. B. Dinh, N. Vo, and G. Medioni. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments. In CVPR, 2011.

[20] F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In ECCV, 2012.

[21] B. Liu, J. Huang, L. Yang, and C. Kulikowsk. Robust Tracking using Local Sparse Appearance Model and K-Selection. In CVPR, 2011.

[22] L. Sevilla-Lara and E. Learned-Miller. Distribution Fields for Tracking. In CVPR, 2012.

[23] C. Bao, Y. Wu, H. Ling, and H. Ji. Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach. In CVPR, 2012.

[24] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust Visual Tracking via Multi-task Sparse Learning. In CVPR, 2012.

[25] Z. Kalal, J. Matas, and K. Mikolajczyk. P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In CVPR, 2010.