HOW VIDEO OBJECT TRACKING IS AFFECTED BY IN-CAPTURE DISTORTIONS?

Roger Gomez Nieto, Hernan Dario Benitez Restrepo, Ivan Cabezas

{roger.gomez,hbenitez}@javerianacali.edu.co; imcabezas@usbcali.edu.co
Pontificia Universidad Javeriana - Cali
Universidad de San Buenaventura- Cali

ABSTRACT

Video Object Tracking -VOT- in realistic scenarios is a difficult task. Image factors such as occlusion, clutter, confusion, object shape, and zooming, among others, have an impact on video tracker methods performance. While these conditions do affect trackers performance, there is not a clear distinction between the scene content challenges like occlusion and clutter, against challenges due to distortions generated by capture, compression, processing, and transmission of videos. This paper is concerned with the latter interpretation of quality as it affects VOT performance. The contribution of this paper is two-fold. We have constructed a database of 537 surveillance videos containing different levels of authentic distortions such as low exposure and out-of-focus. It is available at `https://tinyurl.com/DSVD-Test`. Based on this database, we assessed seven state-of-the-art trackers with the A-R plot performance measure. We demonstrate that in-capture distortions severely hamper VOT methods performance in a non intuitive way.

Index Terms— Video Object Tracking, Video Quality Assessment

1. INTRODUCTION

VOT is a well studied and fast-advancing field. It remains a challenging task since only the initial state of the target is available. Despite the plethora of VOT methods existing in the literature, there is a lack of a detailed study analyzing performance on videos with authentic in-capture and post-capture distortions. Such a study requires a database with videos containing distortions mentioned above in a controlled and quantifiable way. In [1] the authors proposed a standard set of evaluation measures for VOT 2017 [2, 3]. However,this dataset lacks of sequences with videos including in-capture and post-capture distortions in outdoor or indoor environments. A significant number of video quality databases have been designed in the recent years [4–8]. These databases have been generated by systematically distorting, in a controlled manner, a small set of high-quality videos. In fact, most of the existing VOT and Video Quality Assessment -VQA- datasets do not contain simultaneously in-capture and post-capture distortions or only have a single distortion type [9]. Furthermore, these databases do not include authentic in-capture distortions [10].

Deepti et al presented in [11] a video database containing in-capture distortions for VQA. It comprises a total of 208 videos captured using eight different smart-phones. The videos in this database contain six common in-capture distortions such as artifacts, color, exposure, out-of-focus, sharpness, and stabilization. For instance, Tsifouti et al. [12] generated degraded datasets that allow testing how video compression and frame rate reduction affects the performance of video-analytic systems. In this way, they were able to report, an increased false positive ratio due to compression methods. They concluded that performance depends on the specific implementation of the compression software used, on the target bit rate, and on the frame rate. In spite of these advances, to the extent of our knowledge, very little work has been done on the construction databases affected by in-capture distortions for video surveillance applications.

This paper introduces a distorted video surveillance dataset affected by in-capture distortions acquired by four different surveillance cameras and analyzes the impact of real-world in-capture distortions on state-of-the-art VOT methods. The introduced video dataset is suited for testing VOT methods in a varied content of indoor and outdoor scenes of interest to test tracking algorithms. The paper is structured as follows. The set of analyzed trackers are described in Section 2. Applied VOT evaluation metrics along with the introduced authentically distorted video dataset are described in Sections 3 and 4, respectively. Conducted analysis and obtained results are shown in Section 5. Conclusions are stated in Section 6.

2. VIDEO OBJECT TRACKERS

A high performance on the VOT 2017 [1] and VOT 2018 [14] challenges was considered as the criterion for choosing the

The authors acknowledge the funding provided by COLCIENCIAS and Pontificia Universidad Javeriana with the project *Vigilancia Inteligente para la red de cámaras de la Policía Metropolitana de Cali*. The authors would like to thank NVIDIA Corporation for the donation of a TITAN XP GPU used in these experiments. The authors would also like to acknowledge the grant provided by *Comision Fulbright Colombia* to fund the Visiting Scholar Schorlarship granted to Hernan Benitez

analyzed methods.

C-COT tracker learns a discriminative continuous convolution operator as its tracking model [13]. It poses the learning problem in the continuous spatial domain. This enables a natural and efficient fusion of multi-resolution feature maps, e.g. when using several convolutional layers from a pre-trained CNN. The continuous formulation also enables highly accurate localization by sub-pixel refinement [14].

Multi-Cue Correlation Filters for Robust Visual Tracking (MCCT) [15] combines different types of features. It constructs multiple experts through Discriminative Correlation Filter -DCF- tracking the target independently, in each frame. The divergence of multiple experts reveals the reliability of the current tracking, which is quantified for adaptively updating the experts and keep them from corruption. For estimating target scale, MCCT follows the DCCT tracker. The expert with the highest robustness score is selected after evaluating the overall reliability of each node [1].

Efficient Convolution Operators for Tracking -ECO- [16], improves both speed and performance by introducing several efficient strategies. ECO addresses the problems of computational complexity and over-fitting in state-of-the-art DCF trackers by introducing: (i) a factorized convolution operator, which drastically reduces the number of parameters in the model; (ii) a compact generative model of the training sample distribution, that significantly reduces memory and time complexity, while providing better diversity of samples; (iii) a conservative model update strategy with improved robustness and reduced complexity [1].

The Discriminative Scale Space Tracker -DSST- [17] extends the Minimum Output Sum of Squared Errors -MOSSE- tracker [18] with robust scale estimation. The DSST additionally learns a one-dimensional discriminative scale filter, that is used to estimate the target size. For the translation filter, the intensity features employed in the MOSSE tracker is combined with a pixel-dense representation of HOG-features [1].

DeepSTRCF implements a variant of STRCF tracker [19] with deep CNN features. STRCF addresses the computational inefficiency problem of SRDCF tracker from two aspects: (i) a temporal regularization term to remove the need of formulation on large training sets, and (ii) an ADMM algorithm to solve the STRCF model efficiently. Therefore, it can provide more robust models and much faster solutions than SRDCF thanks to online Passive-Aggressive learning and ADMM solver, respectively [14]. This tracker was implemented on MatLab running on a GPU.

LADCF utilizes adaptive spatial regularizer to train low-dimensional discriminative correlation filters [20]. A low-dimensional discriminative manifold space is designed by exploiting temporal consistency, which carries out reliable and flexible temporal information compression, alleviating filter degeneration and preserving appearance diversity. Adaptive spatial regularization and temporal consistency are combined in an objective function, which is optimized by the augmented Lagrangian method. Robustness is further considered by integrating HOG, Colour Names, and ResNet-50 features. For ResNet-50 features, data augmentation [21] is adopted using flip, rotation and blur. This tracker was implemented on MatLab running on a CPU [14].

VITAL [22] carries out tracking using adversarial learning. It uses a generative network to randomly generate masks for augmenting positive samples. Masks are applied to adaptively dropout input features and capture a variety of appearance changes. With the use of adversarial learning, VITAL network identifies the mask that maintains the most robust features of the target objects over a long temporal span. It also, proposes a high-order cost sensitive loss, decreasing the effect of easy negative samples and facilitating the training of classification network. In this way, class imbalance issues, is handled.

3. EVALUATION MEASURES

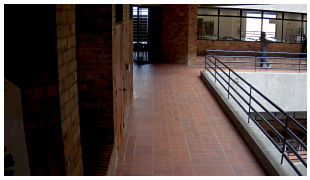
Trackers parameters were set to their respective default values and kept constant during experimentation. Each tracker was executed 30 times on each sequence, considering stochastic processes. This number of executions is enough for statistical evaluation of correlation across the measures. We used the performance measures proposed by [23], for analyzing accuracy and robustness. These are described as follows.

3.1. Average Overlap

The average overlap measure is the most appropriate to be used for tracker comparison. It offers several advantages: simple computation, scale and threshold invariance, exploits the entire sequence, and a clear and concise interpretation. According to the correlation made, the least correlated measures are failure rate and average overlap on re-initialized trajectories. The average overlap measure can be considered as the best choice for measuring the accuracy of a tracker since it takes into account the size of the object and does not require a threshold parameter.

3.2. Failure Rate

The failure rate measure addresses the problem of the VOT length measure. It casts the VOT problem as a supervised system in which a human operator reinitialize the tracker once it fails. The number of required manual interventions per frame is recorded and used as a comparative score. We declare a failure when the bounding box overlap is 0 since we are only interested in the most apparent failure without overlap between regions. The robustness is defined as an exponential failure distribution, $R_s = e^{SM}$. The value of M denotes meantime-between-failures, i.e., $M = \frac{F_0}{N}$, where N is the length of the sequence. The reliability of a tracker can be interpreted as a



(a) Pristine Indoor

(b) Distorted Indoor

(c) Pristine Outdoor

(d) Distorted Outdoor

Fig. 1: Examples of pristine and distorted images within indoor and outdoor environments.

probability that the tracker will still successfully track the object up to S frames since the last failure, assuming a uniform failure distribution that does not depend on previous failures. In this study, we assumed S=30.

4. AUTHENTICALLY DISTORTED VIDEO DATASET

We created a distorted video surveillance dataset of 537 videos affected by in-capture distortions, acquired by four different surveillance cameras [24]. It is available at `https://tinyurl.com/DSVD-Dataset`. The 15 videos and the ground truth bounding boxes with those we perform the tests are available on `https://tinyurl.com/DSVD-Test`. The distortions are out-of-focus, exposure time, and exposure concurrently with out-of-focus. The videos in this dataset have an equal rate I/P frames: 10 fps. This frame rate is typical in commercial applications of video surveillance. The minimization of storage costs also motivates this frame rate selection. The frame size is FHD (1920$\times$ 1080), the color space is three RGB channels and the exposure variation range is $\left\{\frac{1}{480}, \frac{1}{120}\right\}$ seconds. The video dataset also contains H.264/AVC compression post-capture distortions at three different bitrates, resulting in three mirrored video sequences, that change only in the level of compression. The three different bitrates (4700, 1800 and 1200 kbps) were chosen in order to generate degradation all over the distortion scale (from imperceptible to very annoying).

5. RESULTS AND ANALYSIS

The trackers were tested in 15 scenes that contain in-capture distortions such as lack of exposure, out-of-focus and out-of-focus concurrently with lack of exposure in indoor and outdoor environments. Figure 1 presents examples of these test

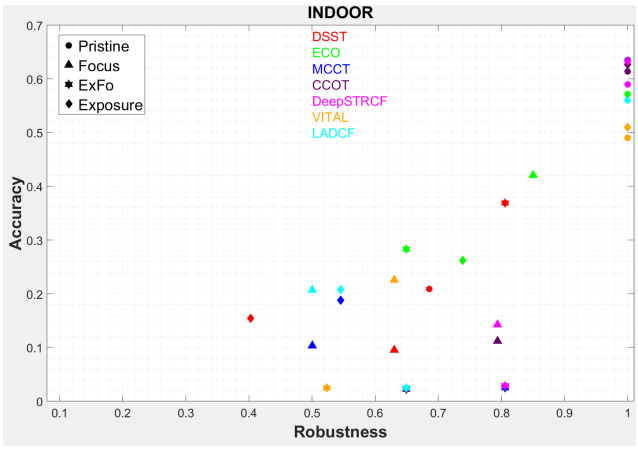


Fig. 2: A-R plot for VOT in an indoor environment with pristine and distorted videos with the same activity.

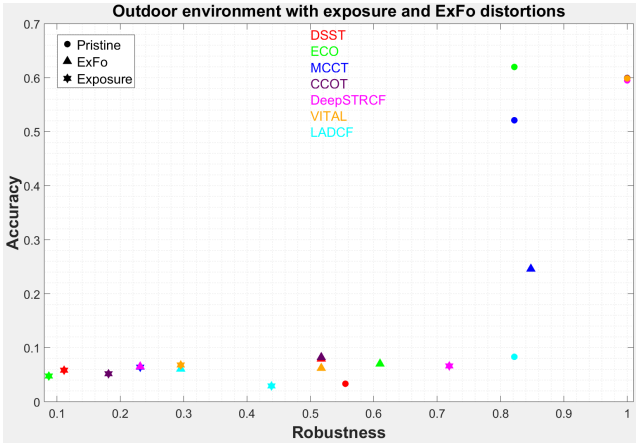


Fig. 3: A-R plot for VOT in and outdoor environment with pristine and distorted videos with the same activity.

sequences, where the difference in image quality between the pristine and distorted image can be identified. It can be noticed that even for the human observer it is difficult to distinguish the people or objects that intervene in the scene in the distorted images.

Figures 2 and 3 show the results of trackers in indoor and outdoor environments. All the videos contain the same activity and were recorded with the same cameras in the same physical space. The only changing aspect is the distortion type. It can be seen that in both environments (indoor and outdoor) the best result is achieved with the pristine videos. Nonetheless, the outdoor environment is more challenging for trackers, possibly due to changes in illumination and scene depth. In both environments, the distortion that most severely affected the trackers was exposure time. Furthermore, in the outdoor environment, the accuracy decreased severely and consistently in all distortions and pristine videos. The tracker who obtained the best results in the indoor environment was

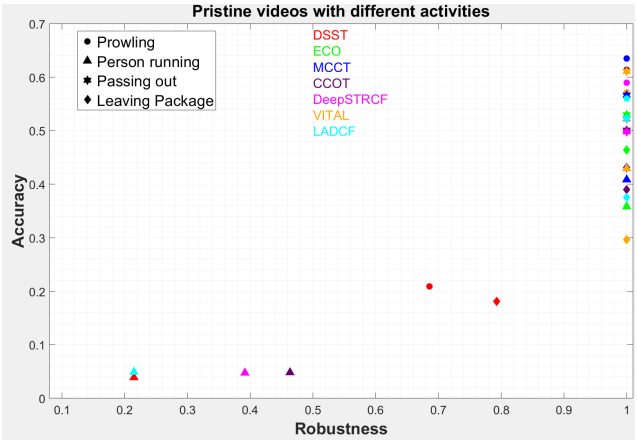


Fig. 4: A-R plot for VOT within an indoor environment in pristine videos (same conditions for all) and different activity.

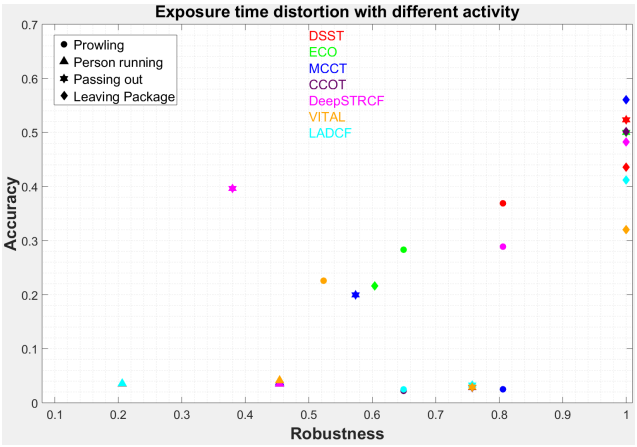


Fig. 5: A-R plot for VOT with exposure time distortion in the same level and different activity in video.

DeepSTRCF. In the outdoor environment, the most accurate tracker was MCCT, and the most robust tracker was DeepSTRCF. We tested the trackers with pristine and distorted videos that contain different activities, to evaluate VOT in realistic scenarios.

In the first scenario, we tested similar activities, now with pristine videos, as is shown in Figure 4. These results demonstrated that in pristine videos, the visual content does not affect the tracker robustness severely, whereas in videos with distortions, the tracker performance (accuracy and robustness) changes in a significant way and have a high dependence on visual content. In the second scenario, we tested four activities (prowling, leaving package person, a person running, a person passing out) on videos from the same camera, with exposure distortion, as is shown in Figure 5. In general, the most challenging video for the trackers was a running person, possibly due to fast changes in object position and the low FPS used (10 FPS). By taking into account the overall performance in all 15 scenes containing all distortions, environments, and activity, the most accurate tracker was DeepSTRCF, and the most robust was VITAL. However, these performances are far from the performance with pristine videos. It highlights the necessity of future VOT methods performance improvement on authentic distortions.

6. CONCLUSIONS

We carried out an analysis of seven state-of-the-art trackers highly ranked in the 2017 and the 2018 VOT challenges [1, 14]. The most innovative aspect of the presented analysis is based on the database used. This paper introduces the Distorted Video Surveillance Database -DVSD. It involves videos affected by in-capture distortions produced by exposure time and out-of-focus variations in challenging indoors and outdoors scenarios. DVSD contains real-world surveillance scenes such as people walking alone, meeting, fighting, passing out, leaving a package in a public place, prowling, and being robbed. In this way, DVSD can be seen as a solid starting point to study the influence of distortions on video tracker performance.

This study concludes that in-capture distortions severely affect the performance of state-of-the-art trackers. As expected, the trackers had the best performance in the pristine videos. Beyond that, the results reflect a poor performance of the trackers due to distortions such as underexposure and out-of-focus. In practice, no specific type of distortion consistently generated the worst performance in all scenes, neither affected all trackers in the same way.

Hence, the design and construction of a robust tracker for these distortions remains as an open question. We believe it can can be answered by creating algorithms relying on perceptual features to compensate the impairments produced by these distortions.

7. REFERENCES

[1] M. Kristan et al., "The Visual Object Tracking VOT2017 Challenge Results," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1949–1972, 2017.

[2] M. Kristan and R. Pflugfelder et al., "The visual object tracking VOT 2014 challenge results," *ECCV*, pp. 1–27, 2014.

[3] M. Kristan and J. Matas et al., "The Visual Object Tracking VOT2015 Challenge Results," in *Proceedings of the IEEE International Conference on Computer Vision*. dec 2015, vol. 2015-Febru, pp. 564–586, IEEE.

[4] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture

quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[5] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, and Tero Vuori, "CVD2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, jul 2016.

[6] A. K. Moorthy I. Katsavounidis A. Aaron C. G. Bampis, Z. Li and A. C. Bovik, "Live netflix video quality of experience database," Accesed on May 2017.

[7] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Moorthy, and Prasanjit Panda, "Subjective and objective quality assessment of mobile videos with in-capture distortions," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1393–1397.

[8] Deepti Ghadiyaram, Alan C Bovik, Hojatollah Yeganeh, and Roman Kordasiewicz, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 989–993.

[9] Kalpana Seshadrinathan, Rajiv Soundararajan, and Alan Conrad Bovik et al., "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, jun 2010.

[10] C. Schuldt and I. Laptev et al., "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* 2004, IEEE.

[11] D. Ghadiyaram, J. Pan, and A. C. Bovik et al., "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[12] Anastasia Tsifouti and Moustafa M. Nasralla et al., "A methodology to evaluate the effect of video compression on the performance of analytics systems," in *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, Colin Lewis and Douglas Burgess, Eds. oct 2012, SPIE.

[13] M. Danelljan and A. Robinson et al., "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 472–488, Springer International Publishing.

[14] Matej Kristan and Ales Leonardis et al., "The sixth visual object tracking vot2018 challenge results," 2018.

[15] N. Wang and W. Zhou et al., "Multi-cue correlation filters for robust visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4844–4853.

[16] M. Danelljan and G. Bhat et al., "Eco: Efficient convolution operators for tracking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] M. Danelljan, G. Hger, and F. S. Khan et al., "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, Aug 2017.

[18] D. Bolme and J. R. Beveridge et al., "Visual object tracking using adaptive correlation filters," *Computer Vision and Pattern Recognition CVPR 2010 IEEE Conference on*, pp. 2544–2550, 2010.

[19] F. Li and C. Tian et al., "Learning spatial-temporal regularized correlation filters for visual tracking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] Tianyang Xu and Zhen-Hua Feng et al., "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking," *https://arxiv.org/abs/1807.11348*, 2018.

[21] G. Bhat and J. Johnander et al., "Unveiling the power of deep tracking," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[22] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, and Wangmeng Zuo, "Vital: Visual tracking via adversarial learning," *Arxiv*, vol. https://arxiv.org/abs/1804.04273, 2018.

[23] L Cehovin, A Leonardis, and M Kristan, "Visual Object Tracking Performance Measures Revisited," *Image Processing, IEEE Transactions on*, vol. 25, no. 3, pp. 1261–1274, 2016.

[24] Roger Gomez Nieto, H. D. Benitez-Restrepo, and Ivan Mauricio Cabezas, "Evaluation of object trackers in distorted surveillance videos," *Arxiv*, vol. http://arxiv.org/abs/1804.01624v1, 2018.