Preserving Temporal Consistency in Videos Through Adaptive SLIC

Han Zhang[1], Riaz Ali[2], Bin Sheng[2], Ping Li[3], Jinman Kim[4], and Jihong Wang[5]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, People's Republic of China
[2] Shanghai Jiao Tong University, Shanghai, People's Republic of China
shengbin@sjtu.edu.cn
[3] The Hong Kong Polytechnic University, Hong Kong, People's Republic of China
[4] The University of Sydney, Sydney, Australia
[5] Shanghai University of Sport, Shanghai, People's Republic of China
wjh@sus.edu.cn

Abstract. The application of image processing techniques to individual frames of video often results in temporal inconsistency. Conventional approaches used for preserving the temporal consistency in videos have shortcomings as they are used for only particular jobs. Our work presents a multipurpose video temporal consistency preservation method that utilizes an adaptive simple linear iterative clustering (SLIC) algorithm. First, we locate the inter-frame correspondent pixels through the SIFT Flow and use them to find the respective regions. Then, we apply a multiframe matching statistical method to get the spatially or temporally correspondent frames. Besides, we devise a least-squares energy-based flickering-removing objective function by taking into account the inter-frame temporal consistency and inter-region spatial consistency jointly. The obtained results demonstrate the potential of the proposed method.

Keywords: Video processing · Adaptive SLIC · Temporal consistency

1 Introduction

Maintaining the temporal consistency is an essential task in video processing because the temporal consistency is one of the essential video features and has been used in different applications [15,16,18,22]. Temporal inconsistency results in artifacts, like unnatural inter-frame tonal changes or brightness fluctuations, which decrease the video quality significantly [11,23]. Although the flickers may not be easily observed when adjacent video frames are seen individually, they will be apparent when the video is played. Also, these artifacts adversely affect video matching tasks such as motion estimation [8,9]. Thus, preserving the video consistency is a crucial yet laborious problem in video processing.

H. Zhang and R. Ali contributed equally to this work.

Some studies, like and [2], solve the temporal consistency problem in the form of energy minimization. Nevertheless, their applications are limited to removing flickers in intrinsic video decomposition. Lang et al. [10] eliminate the flickers through edge-aware filter employed in the temporal domain, but their method only deals with high-frequency perturbations. The technique of [7] removes flickers from video halftoning by using an error diffusion of temporal and spatial terms. The work of Dong et al. [5] uses non-flickering frames to rebuild the flickering frames. However, it is suited to eliminating flickers generated due to directly employing the image enhancement algorithm on a video. In [3], a frame is rebuilt from a neighboring frame to retain the inter-frame temporal consistency. However, it is not feasible because a video may not contain the same object continuously in adjacent frames [17]. Some authors have proposed compensation-based techniques that aim to eliminate the artifacts by aligning the inter-frame brightness or tonal level. In [19], the atmospheric light values are determined with the aid of the adaptive temporal average to eliminate the flickering effects. Farbman et al. [6] and Wang et al. [21] align the video frames by a specified number of designated *key* frames. After interpolating the transformation between the model video and the input video, the authors of [1] enhance the color grade by hand by choosing some key frames that can depict the transformation curve. However, these methods are limited because of requiring to select the key frames first.

In this paper, we propose a multipurpose flickering-removal and spatiotemporal consistency preservation technique. We develop an adaptive SLIC algorithm that creates superpixels from every frame. We also propose a multiframe matching statistical model to capture the frame that is correspondent to other frames temporally or spatially. First, our method matches the inter-frame corresponding pixels through the SIFT Flow [12] algorithm, and then, calculates the corresponding regions using those pixels. Lastly, we use the inter-frame corresponding regions and frame interval to match the corresponding frames. Because several studies have restored temporal coherence in videos with the help of least squares energy [14], we develop the objective function of temporal consistency in the form of least-squares energy comprising temporal and spatial consistencies terms. The former term preserves the inter-frame tonal or illumination variations consistency, and the latter term maintains the consistency in adjacent regions' difference of changes.

2 Methodology

Figure 1 illustrates the overview of our proposed technique. The following sections explain each of the steps.

Adaptive SLIC: The conventional SLIC manually computes the number of superpixels in a repetitive manner, which is a tedious task. We develop an adaptive SLIC algorithm that automatically produces the number of superpixels. We convert the RGB color to HSV color space because, for small color ranges, HSV

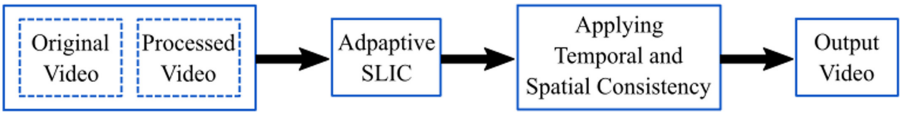

Fig. 1. Overview of the proposed method.

is perceptually uniform. In our experiments, we use the unequal interval to quantize the H, S, V values, and the average occurrence number to decide the number of superpixels.

Applying Temporal Consistency: We reproduce an output frame, denoted as P_x, from its temporally matching frames $f(P_x)$ to preserve the inter-frame coherence. Suppose P_x^k denotes the k^{th} iteration's output frame, then the temporal consistency $F_t(P_x^k)$ can be calculated as Eq. (1):

$$\begin{aligned} F_t(P_x^k) = \mu(k) \cdot \sum_{I_m \in f^q(I_x)} \psi_t(I_x, I_m) \left\| P_x^k - \text{warp}(P_m^{\alpha(k)}) \right\|^2 \\ + \nu(k) \cdot \sum_{I_m \in f^s(I_x)} \psi_t(I_x, I_m) \left\| P_x^k - \text{warp}(P_m^{\beta(k)}) \right\|^2 \end{aligned} \tag{1}$$

where I_x, $f(I_x)$, respectively, denote the actual frame and the set of its corresponding frames, I_m is the matching frame of the actual frame, $f^q(I_x)$ and $f^s(I_x)$ represent the sets of previous and subsequent matching frames of I_x, respectively. $\psi_t(I_x, I_m)$ denotes the temporal consistency weight, and the warped output frame from P_m^k is represented as $\text{warp}(P_m^k)$, where warp() is a function that uses the optical flow [13] to recreate a resultant frame from its matching frame.

Applying Spatial Consistency: We calculate an output frame's spatial consistency in the k^{th} iteration, as shown in Eq. (2):

$$F_s(P_x^k) = \sum_{a=1}^{A_x} \sum_{\Upsilon_x^b \in \Omega(\Upsilon_x^a)} \psi_s(\Upsilon_x^a, \Upsilon_x^b) \left\| P_x^k - \text{warp}(P_\xi^{\Gamma(x,\xi,k)}) \right\|^2 \tag{2}$$

where A_x denotes the count of regions inside I_x. Υ_x^a and $\Omega(\Upsilon_x^a)$, respectively, represent a region and a set of its all adjacent regions. To preserve the inter-region spatial consistency, we decrease the change between the output frame and its respective spatially correspondent frame. Υ_x^b is an adjacent region of Υ_x^a, and the regions with the most matching pixels to Υ_x^b is denoted as $\hat{\Upsilon}_x^b$. I_ξ represents the frame containing the $\hat{\Upsilon}_x^b$. We retrieve P_x through warping its spatially matching output frame P_ξ. Therefore, to preserve the spatial consistency, we maintain the consistency in the adjacent regions' variation difference.

Combined Optimization: We optimize the output frame P_x by merging the temporal and spatial consistencies. Equation (3) shows the objective function comprising both the consistency terms

$$\arg\min_{P_x^k} \int [F_g(P_x^k) + \eta_1 F_t(P_x^k) + \eta_2 F_s(P_x^k)] \mathrm{d}u \tag{3}$$

where u denotes the spatial location in I_x, η_1, and η_2 are, respectively, the weight functions of temporal consistency and spatial consistency.

3 Experimental Results

We have performed experiments on two datasets, SegTrack [20] and Chen and Corso [4]. Figure 2(a)–(c) shows the visual results of flickering removal by our method. The girl's neck and cheek region are darker in the processed video. It is evident that our method satisfactorily removes these effects.

Fig. 2. The flickering removal results on the two frames of the CC video. (a)–(c) actual frame, processed frame, and the result of our approach, respectively.

4 Conclusion

In this paper, we present our proposed method of preserving temporal consistency in videos using adaptive SLIC. We employ the consistency to maintain the tonal shifts or illumination fluctuations constant among the adjacent regions. We find the regions through a new adaptive SLIC segmentation algorithm that uses the color details to automatically calculate the count of superpixels. The proposed temporal consistency solution outperforms previous techniques as our warping procedure comprises both; the spatial and temporal consistencies. The results obtained during the experiments demonstrate that our method provides satisfactory performance in video flickering-removal.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFF0300903, in part by the National Natural Science Foundation of China under Grant 61872241 and Grant 61572316, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 15490503200, Grant 18410750700, Grant 17411952600, and Grant 16DZ0501100.

References

1. Bonneel, N., Sunkavalli, K., Paris, S., Pfister, H.: Example-based video color grading. ACM Trans. Graph. **32**(4), 39:1–39:12 (2013)
2. Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S., Pfister, H.: Interactive intrinsic video editing. ACM Trans. Graph. **33**(6), 197:1–197:10 (2014)
3. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. ACM Trans. Graph. **34**(6), 196:1–196:9 (2015)
4. Chen, A.Y.C., Corso, J.J.: Propagating multi-class pixel labels throughout video frames. In: Western New York Image Processing Workshop, pp. 14–17 (2010)
5. Dong, X., Bonev, B., Zhu, Y., Yuille, A.L.: Region-based temporally consistent video post-processing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722 (2015)
6. Farbman, Z., Lischinski, D.: Tonal stabilization of video. ACM Trans. Graph. **30**(4), 89:1–89:10 (2011)
7. Hsu, C.Y., Lu, C.S., Pei, S.C.: Video halftoning preserving temporal consistency. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 1938–1941 (2007)
8. Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D.: Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Trans. Syst. Man Cybern. Syst. **49**(9), 1806–1819 (2019)
9. Karambakhsh, A., Kamel, A., Sheng, B., Li, P., Yang, P., Feng, D.D.: Deep gesture interaction for augmented anatomy learning. Int. J. Inf. Manag. **45**, 328–336 (2019). https://doi.org/10.1016/j.ijinfomgt.2018.03.004. http://www.sciencedirect.com/science/article/pii/S0268401217308678
10. Lang, M., Wang, O., Aydin, T., Smolic, A., Gross, M.: Practical temporal consistency for image-based graphics applications. ACM Trans. Graph. **31**(4), 34:1–34:8 (2012)
11. Li, C., Chen, Z., Sheng, B., Li, P., He, G.: Video flickering removal using temporal reconstruction optimization. Multimedia Tools Appl. **79**, 4661–4679 (2019)
12. Liu, C., Yuen, J., Torralba, A.: SIFT flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011)
13. Liu, C.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (2009)
14. Mantiuk, R., Daly, S., Kerofsky, L.: Display adaptive tone mapping. ACM Trans. Graph. **27**(3), 1–10 (2008)
15. Meng, X., et al.: A video information driven football recommendation system. Comput. Electr. Eng. **85**, 106699 (2020). https://doi.org/10.1016/j.compeleceng.2020.106699
16. Müller, M., Zilly, F., Riechert, C., Kauff, P.: Spatio-temporal consistent depth maps from multi-view video. In: 2011 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp. 1–4 (2011)

17. Reso, M., Jachalsky, J., Rosenhahn, B., Ostermann, J.: Occlusion-aware method for temporally consistent superpixels. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1441–1454 (2019)
18. Sheng, B., Li, P., Zhang, Y., Mao, L.: GreenSea: visual soccer analysis using broad learning system. IEEE Trans. Cybern. 1–15 (2020). https://doi.org/10.1109/TCYB.2020.2988792
19. Shin, D.K., Kim, Y.M., Park, K.T., Lee, D.S., Choi, W., Moon, Y.S.: Video dehazing without flicker artifacts using adaptive temporal average. In: The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014), pp. 1–2 (2014)
20. Tsai, D., Flagg, M., Nakazawa, A., Rehg, J.M.: Motion coherent tracking using multi-label MRF optimization. Int. J. Comput. Vision **100**(2), 190–202 (2012)
21. Wang, C.M., Huang, Y.H., Huang, M.L.: An effective algorithm for image sequence color transfer. Math. Comput. Model. **44**, 608–627 (2006)
22. Wang, Z., Chen, X., Zou, D.: Copy and paste: temporally consistent stereoscopic video blending. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 3053–3065 (2018)
23. Zhang, P., Zheng, L., Jiang, Y., Mao, L., Li, Z., Sheng, B.: Tracking soccer players using spatio-temporal context learning under multiple views. Multimedia Tools Appl. **77**(15), 18935–18955 (2017). https://doi.org/10.1007/s11042-017-5316-3