Quality Enhancement for Feature Matching on Car Black Box Videos

Christian Simon, Man Hee Lee, and In Kyu Park
Department of Information and Communication Engineering, Inha University
Incheon 402-751, Korea
{cssen@inha.edu, maninara@hotmail.com, pik@inha.ac.kr}

***Abstract*—Video has difficulty to maintain consistent intensity and color tone from frame to frame. Particularly, it happens when imaging device such as black box camera has to deal with fast changing illumination environment. However, conventional automatic white balance algorithms cannot handle this good enough to maintain tone consistency, which is observed in most commercial black box products. In this paper, a novel tone stabilization is proposed to enhance the performance of further applied algorithms like detecting and matching visual features across video frames. The proposed technique utilizes multiple anchor frames as references to smooth tone fluctuation between them. Experimental result shows the improvement of tone consistency as well as feature detection and matching accuracy on car black box videos with varying tone over time.**

I. Introduction

As mobile imaging devices become more and more popular, we can see more unwanted images or videos taken in an improper way. The problem often happens especially for the automatic device which cannot be controlled manually because the device itself is originally intended to use easily without user intervention. For example, some devices, such as black box inside a car, are not equipped with many settings because it is not expected to adjust the device while driving.

Car black box is facilitated with automatic white balance which is able to adjust tone automatically in camera system. In term of the quality for human vision, this feature is very helpful because it can record the images while reducing dynamic range problem. However, this capability is not yet satisfactory in term of applying computer vision algorithms such as object matching and recognition. It often happens that, because the tone is various and it can still bother the matching process among frames because of the changes.

Black box video is a helpful evidence for some occasions such as car accident and street viewing. Recent drivers usually equip a black box with their car. Commercial black boxes in the market have high resolution camera and several gigabytes of memory, which can record a few hours of high definition video. Because of this, technical demand is overwhelmed to conduct image and video analysis to extract meaningful information. However, since automatic video analysis is not a trivial thing to be solved, it needs sophisticated computer vision algorithms to be applied. For instance, robust images matching is quite popular to make relation from one frame to another which share same features. Moreover, the geometric and photometric features of black box camera lens can be far from standard one to the specific purpose such as 3D reconstruction or image understanding.

Several computer vision techniques such as SLAM (Simultaneous Localization and Mapping) utilize multiple images from video. This technique employs feature detection and matching. Some algorithms have been proposed for robust feature detection and description such as SIFT (Scale-Invariant Feature Transform) [6] and SURF (Speeded-Up Robust Features) [1]. However, the localization and scale/orientation estimation of detected features could be inconsistent under different tone condition.

In this paper, we proposed a technique for stabilizing tone to enhance the performance of further applied algorithms like detecting, matching, and tracking image features across frames of car black box video. The proposed technique for tone stabilization of video input uses multiple anchor frames as references to smooth tone changes over time. Furthermore, color normalization is applied for each anchor frame and adjustment map is formed using anchor frames to adjust the tone change in a frame. Furthermore, this paper unfolds an analysis of correlation and effect between tone stabilization and the improvement of feature matching accuracy on car black box videos with fluctuated tone over video frames.

This paper is organized as follows, In Section II, a few related work is introduced. In Section III, the proposed method of quality enhancement is described. Experimental result and comparison is shown in in Section IV. Finally, we give a conclusive remark in Section V.

II. Related Work

A lot of low level algorithms have been proposed to improve the quality of the image frame such as color transfer [9], [15] and contrast enhancement [3], [8]. However, this kind of methods has limitation that they can be applied only for a single image. They produce incoherent result from frame to frame in video even though it is possible to apply the method on frame by frame basis. Some video processing algorithms have also been applied to enhance the quality of video. Wang and Huang [13] proposed a novel color transfer to make coherent frames with three reference frames. This method is not fully automatic color transfer because the user has to choose particular source frames. The previous work which is the most suitable for tonally varying video like black box video was proposed by Farbman and Lischinski [4]. It was

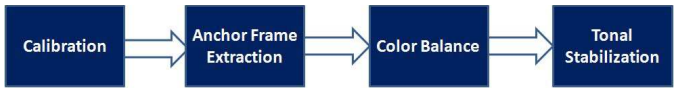


Fig. 1. Overview of the proposed method to enhance black box video.

further improved by Simon and Park [12] such that it chooses anchor frames automatically to adjust the tone of the other frames while the original method in [4] needs user to choose the anchor frames is a manual way.

III. Proposed Method

The recorded video from black box should be improved to produce a better quality of video by considering photometric as well as geometric aspects. Camera calibration is applied to each frame to remove geometric distortion caused by camera lens. Moreover, the proposed method for producing tonally consistent video employs the work in [12] to adjust the massive tone difference between frames automatically. The overview of the proposed method is shown in Fig. 1.

A. Geometric Correction

Frames recorded using black box has serious radial distortion due to the wide angle property of the camera because the device is intended to capture the information in a scene as much as possible. We employ the popular method of Zhang [16] for initialization and Heikkila and Silven's method [5] for refining intrinsic model including two additional distortion coefficients. The geometric calibration is operated only once as an off-line procedure because the intrinsic parameters of black box camera is static.

B. Photometric Correction

1) Anchor Frames Extraction: In order to select anchor frames automatically, it is necessary to cluster a group of ordered frames, so that the candidates of anchor frames are found out of the clusters. In our approach, K-means clustering is employed to this purpose, in which average color vector is used for the feature vector of each frame. The number of clusters, K, is obtained adaptively as follows [7].

$$K = \sqrt{\frac{n}{2}} \tag{1}$$

where n is the total number of frames in a video. The anchor frame of each particular cluster is extracted by selecting the frame with minimum distance to the centroid of the cluster.

2) Color Balance: Each extracted anchor frame may suffer from different tone composition caused by rapidly changing environment or automatic white-balance of imaging system. A color balance method helps the composition of red (R), green (G), and blue (B) channels to maintain their coherency, thus it improved tone consistency among anchor frames. One of the algorithms that can be utilized to maintain the white balance in each anchor frame is gray-world assumption [10]. Gray-world assumption contributes in balancing RGB composition of an image as stated below.

$$\begin{aligned}(R', G', B') &= \\ &= \left(\frac{3\Sigma R}{\Sigma R + \Sigma G + \Sigma B}, \frac{3\Sigma G}{\Sigma R + \Sigma G + \Sigma B}, \frac{3\Sigma B}{\Sigma R + \Sigma G + \Sigma B}\right)\end{aligned} \tag{2}$$

where R', G', and B' are the normalized ratios for R, G, and B channels which can be applied to normalize each pixel by dividing it into corresponding channel $(\frac{R}{R'}, \frac{G}{G'}, \frac{B}{B'})$. Then, all anchor frames have relatively similar distribution of color values.

3) Tone Adjustment on Frames: The proposed method of tone adjustment is motivated by Farbman and Lischinski's work [4]. The adjustment map, A'_{i+1}, which is a map containing the difference between consecutive frames, f_i and f_{i+1}, in *Lab* color space is employed. Small size bilateral filter (5 $\times$ 5) is applied to reduce noise beforehand.

In order to obtain adjustment map from f_i to f_{i+1}, we need to calculate the initial adjustment map, A_{i+1}, and the robust map, R_{i+1}, as follows.

$$A_{i+1}(x) = \begin{cases} A_i(x) + (f_i(x) - f_{i+1}(x)) & \text{for each } x \,\epsilon\, R_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where R_{i+1} is a robust map to make a mask to handle large differences around moving object boundaries defined as

$$R_{i+1} = \{x \,|\, (L_i(x) - \bar{L}_i) - (L_{i+1}(x) - \bar{L}_{i+1}) \,<\, \tau\}. \tag{4}$$

In (4), L_i denotes the luminance of the current frame and $\bar{L}_i$ defines the average luminance of the current frame. In robust map R_{i+1}, there are regions which are not covered because they have difference more than or equal to the threshold τ (0.05 in our implementation).

Using the initial adjust map, the final adjustment map, A'_{i+1}, is calculated as follows.

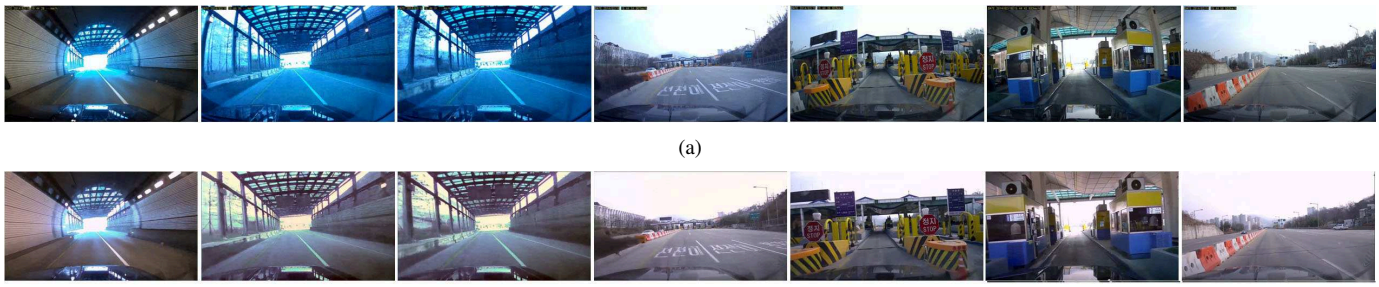
$$A'_{i+1}(x) = \frac{\Sigma G(x, x_r) A_{i+1}(x_r)}{\Sigma \chi(x, x_r) A_{i+1}(x_r)} \tag{5}$$

in which G denotes Gaussian weight function with center pixel at x and its surrounding pixels at x_r. $\chi(\cdot)$ is zero if the pixel value within A_{i+1} is zero, while it is one otherwise.

Lastly, to produce stabilized tone among frames using the adjustment map, it is necessary to transform frames from *RGB* to *Lab* color space. Adjustment map between neighboring anchor frames has to be weighted to produce smooth changes between them. Each frame, with index i, has adjustment map weight function defined as follows.

$$w_i = \frac{N_f - i}{N_f} \tag{6}$$

where N_f is the number of frames between two anchor frames. Frames located exactly beside anchor frames are indexed as 0. New adjusted frame can be obtained by $f'_i = f_i + w_i A'_i$. Note that, the adjustment maps for each frame come from two anchor frames and corresponding weight applies linearly.



(a)

(b)

Fig. 2. Original and result video comparison. (a) Frames in the original video. (b) Frames in the tone stabilized video.

C. Feature Detection and Matching

Existing feature description techniques are applied to evaluate the improved quality of the enhanced black box video. The goal is not to compare the performance of different descriptors but to see how it works better on the enhanced video. In our approach, SIFT is used for initial feature localization and the following descriptors are used for description and matching features.

1) LIOP (Local Intensity Order Pattern) [14]: LIOP descriptor is generated using ordinal information, in which local patch is divided into several ordinal bins or sub-regions. Then, these bins are utilized to construct LIOP descriptor by accumulating the relationship score of the neighboring sample points in each ordinal bin. LIOP descriptor is known as a robust technique to varying illumination and image rotation. Descriptor matching is usually performed by comparing nearest neighbor distance ratio.

2) KLT (Kanade-Lucas-Tomasi) [11]: KLT descriptor is applicable if two images do not have large displacement of the local area. KLT relies on image gradient and the method basically minimizes the dissimilarity (ϵ), which is defined as follows.

$$\epsilon = \int\int_W \{J(\mathbf{Ax}+\mathbf{d}) - I_o(\mathbf{x})\}^2 w(\mathbf{x}) d\mathbf{x} \qquad (7)$$

where $\mathbf{A} = \mathbf{I} + \mathbf{D}$, $\mathbf{D}$ and $\mathbf{I}$ are the deformation matrix and 2×2 identity matrix respectively. I_o and J represent the current image and the second image to be matched. Then, $\mathbf{x}$ and $w(\mathbf{x})$ denote image coordinate and weight function, respectively.

3) SIFT (Scale-Invariant Feature Transform) [6]: In SIFT descriptor, maxima and minima in DoG (Difference of Gaussian) scale space are used to localize the keypoints. Briefly, nearest neighbor test and modified k-d tree are utilized for features matching and indexing. SIFT keypoints include location, scale, and orientation for descriptor generation. All matches with Euclidean distance below a threshold (0.8 in our implementation) are classified to outliers.

4) BRIEF (Binary Robust Independent Elementary Feature) [2]: BRIEF uses the characteristic of binary strings to match features without directly searching particular descriptor. Moreover, it should accompany with another detector algorithm. The difference of intensity in a smoothed image patch is utilized as BRIEF descriptor in a binary string form. Hamming distance is the metric to find the distance between descriptors.

IV. Experimental Results

In our experiment, a commercial black box, Itronics ITB-100HD, is used to capture videos with full high definition resolution (1920×1080). It captures 24 FPS videos in ultra-wide viewing angle of 144 degree. The focus and exposure are automatically controlled by the device. Note that in this device many image preprocessing techniques are applied already to improve the video quality. The captured video has severe photometric fluctuation due to the different illumination condition while car moves (*e.g.* entering and exiting a tunnel and changing the direction of car). The video is only captured in daytime since outside scene at night is too dark to process with computer vision algorithms.

The proposed algorithm is implemented on Intel Core i7 3.5 GHz CPU and NVIDIA GeForce GTX 680 GPU. Processing HD image needs a huge running time. Therefore, the adjustment map is processed in the reduced size (in our implementation, height and width are divided by 4). The size is returned back to the original after the final adjustment map is obtained. Furthermore, the implementation is partially parallelized on the GPU, especially the process of producing adjustment map. In this occasion, we can reach the processing time about 1.4 seconds per frame.

The typical result of tone stabilization is shown in Fig. 2. Tone inconsistency is obviously observed in the input video. The tone shown in the sample frames fluctuates due to automatic white balance and the illumination change at outside. As shown in Fig. 2 (b), our method shows significant result to improve inconsistent tone and radial distortion in a black box video. Fig. 3 compares the average of RGB channels in the sequential frames of original and enhanced videos. The comparison shows that the original video has more fluctuation in average of RGB channels rather than fluctuation in the enhanced video. Hence, consistent tone across frames in a video can be verified by measuring fluctuation across frames.

To evaluate and analyze the performance improvement in feature detection and matching, we deliver the result of feature description and matching between consecutive frames of whole video sequence. We test four state-of-the-art feature de-

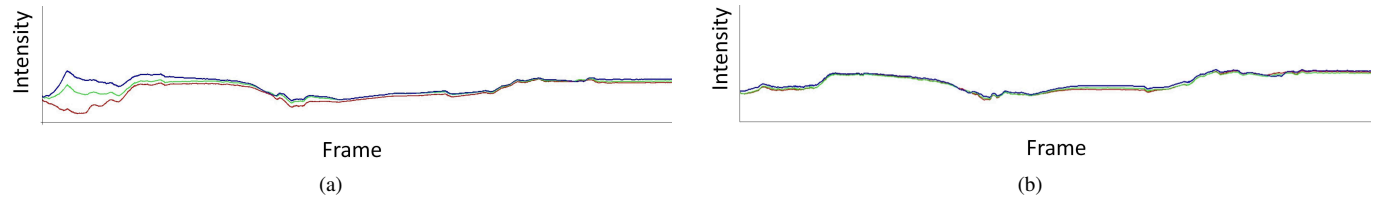


Fig. 3. Histogram comparison of input and enhanced videos in Fig. 2. (a) Histogram composition of the original video. (b) Histogram composition of the tone stabilized video.

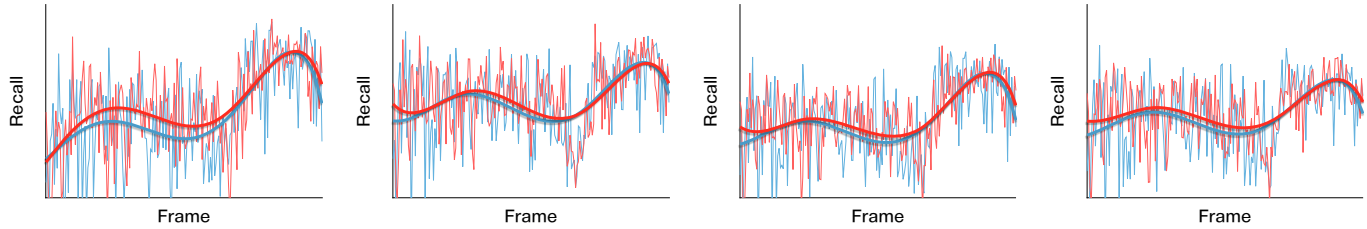


Fig. 4. The recall plot of different descriptors. Blue and red lines denote the recall rate of the original and enhanced video. Thin and thick lines represent the recall plot of each frame and averaged (using ten neighboring frames) frames. From left to right: LIOP [14], KLT [11], SIFT [6], and BRIEF [2].

scriptors, *i.e.* LIOP [14], KLT [11], SIFT [6], and BRIEF [2]. Note that initial key points are detected using SIFT detector. In Fig. 4, it is clearly shown that the tone stabilized video has higher recall rate for all four descriptors. Quantitatively, the recall increases by 3.6% more than the original video in average. Moreover, the tone stabilized video can match inliers 5.8% more than original one. Since the tone variation happens in the first half of the video, the improvement of the matching performance is clearly visible on the first half of the plot. Note that the quantitative improvement is not with a large number of percentage because the large tone variation is not observed in the whole sequence of frames but in small portion of the sequence.

V. Conclusions

In this paper, we proposed an effective method to improve the tone consistency among sequential frames in a block box video. The proposed method is very practical especially as a preprocessing method to help computer vision algorithm such as feature detection and matching perform better. We also presented the experimental proof that the proposed method has the reciprocity between tone stabilized video with the improvement in feature detection and matching. It was shown that that the tonal fluctuation has effect on the accuracy of feature descriptor and matching process. The proposed algorithm significantly improved the performance for various feature descriptors.

Acknowledgment

This work was supported by the IT R&D program of MSIP/ KEIT. [10047078, 3D reconstruction technology development for scene of car accident using multi view black box image].

References

[1] H. Bay, T. T. A. Ess, and L. V. Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.
[2] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *Proc. of European Conference on Computer Vision*, pages 778–792, September 2010.
[3] T. Celik and T. Tjahjadi. Automatic image equalization and contrast enhancement using gaussian mixture modeling. *IEEE Trans. on Image Processing*, 21(1):145–156, January 2012.
[4] Z. Farbman and D. Lischinski. Tonal stabilization of video. In *ACM Trans. on Graphics*, pages 89:1–89:9, July 2011.
[5] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 1106–1112, June 1997.
[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
[7] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1976.
[8] J. Mukherjee and S. K. Mitra. Enhancement of color images by scaling the dct coefficients. *IEEE Trans. on Image Processing*, 17(10):1783–1794, October 2008.
[9] J. Rabin, J. Delon, and Y. Gousseau. Regularization of transportation maps for color and contrast transfer. In *Proc. of IEEE International Conference on Image Processing*, pages 1933–1936, September 2010.
[10] A. Schiele and A. Waibel. Gaze tracking based on face-color. In *Proc. of International Workshop on Automatic Face and Gesture Recognition*, pages 344–349, June 1995.
[11] J. Shi and C. Tomasi. Good features to track. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, June 1994.
[12] C. Simon and I. K. Park. Video tonal stabilization with automatic multiple anchor frames. In *Proc. of IEEE International Symposium on Consumer Electronics*, pages 11–12, June 2014.
[13] C. M. Wang and Y. H. Huang. A novel color transfer algorithm for image sequences. *Information Science and Engineering*, 20(6):1039–1056, June 2004.
[14] Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. In *Proc. of IEEE International Conference on Computer Vision*, pages 603–610, November 2011.
[15] H. Xu, G. Zhai, and X. Yang. No reference measurement of contrast distortion and optimal contrast enhancement. In *Proc. of International Conference on Pattern Recognition*, pages 1981–1984, November 2012.
[16] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. of IEEE International Conference on Computer Vision*, pages 666–673, September 1999.