Super High Dynamic Range Video

Yuka Ogino\textsuperscript{*}, Masayuki Tanaka\textsuperscript{*}, Takashi Shibata\textsuperscript{*}\textsuperscript{†}, and Masatoshi Okutomi\textsuperscript{*}

\textsuperscript{*}Department of Mechanical and Control Engineering
Tokyo Institute of Technology

\textsuperscript{†}Data Science Research Labs, NEC Corporation

Abstract—High dynamic range (HDR) imaging is highly demanded in computer vision algorithms. An HDR image is composed with several low dynamic range (LDR) images, which usually have some disparities. In many HDR imaging algorithms, the disparities are estimated based on the texture information of the LDR images. However, the texture information is often lost completely if scenes include extremely bright and dark regions simultaneously. Recently, super high dynamic range (SHDR) imaging algorithm has been proposed where the disparities are estimated based on the segment shapes instead of the textures for handling such extreme scenes. In this paper, we extend the SHDR imaging algorithm to SHDR video generation introducing temporal smoothness terms. The temporal smoothness terms improve the temporal stability and the precision of the disparity estimation. Quantitative and qualitative evaluations demonstrate that the proposed algorithm outperforms existing algorithms.

I. INTRODUCTION

High dynamic range (HDR) imaging is widely used. Several consumer digital cameras already have HDR imaging functionality. Nevertheless, the improvement of the dynamic range is restricted. A very high dynamic range is often necessary to capture actual scenes which include strong back lighting and/or car headlights. Existing HDR imaging algorithms cannot manage such scenes properly. This point persists as a challenge related to HDR imaging.

An HDR image is reconstructed with multiple low-dynamic-range images taken under different exposure settings. Multiple images are usually taken sequentially by a single camera while changing the exposure setting. Image alignment between different exposure images is necessary to reconstruct an HDR image of a dynamic scene or the scene which includes moving objects. Another approach to take multiple exposure images is using a multiple camera system. Even if multiple cameras are temporarily synchronized, there are disparities among multiple images because the viewpoint of each camera is spatially different. Therefore, image alignment is also required for the HDR imaging with the multiple camera system. Optical-flow-based algorithms have been proposed to reconstruct an HDR image handling the disparity between multiple exposure images [1]–[10]. In the optical-flow-based algorithms, the optical flows between the different exposure images are estimated based on texture similarities. Therefore, the optical-flow-based algorithms implicitly assume that texture information is available from all different exposure images. However, this assumption is not always true, especially for scenes that require a very high dynamic range. In this case, some exposure images include regions in which texture information is completely lost because of saturation or crushed-black. The optical-flow-based algorithms produce severe artifacts for those regions. A weighting approach [11]–[13] is another approach to handle displacement between different exposure images. The weighting approach assigns a lower weight to ghost pixels or dynamic regions to reduce ghost artifacts, where the weight is used to componse the HDR image. However, the weighting approach also has the same problem as the optical-flow-based algorithm because the weight is calculated based on texture similarities.

In order to manage the problems of saturation and crushed-black, Hayami et al. proposed so-called super HDR imaging for such a scene with an extremely wide range of scene radiance [14]. A key idea of their algorithm is to align images based on the shapes of segmented regions instead of texture information, so that they can align extremely underexposed or overexposed regions without being affected by saturation or crushed-black.

HDR video generation is another challenge related to HDR imaging. A specialized HDR camera system that includes multiple cameras and beam splitters has been proposed for the HDR video generation [2]. That HDR camera system can capture images simultaneously under different exposures without disparity. However, such systems are bulky and expensive. In addition, the number of the cameras are usually limited to two or three. Since more than five cameras are usually necessary to capture actual scenes with an extremely wide range of scene radiance, their camera systems cannot be used for such scenes.

In this paper, We propose a super high dynamic range (SHDR) video generation algorithm. Different exposure image sequences are captured by a synchronized multiple camera system. The number of cameras can be increased easily for scenes with an extremely wide range of scene radiance. We extend the SHDR imaging [14] considering temporal smoothness. The proposed SHDR video generation algorithm consists of three steps, i.e. a segmentation step, a segment-based alignment step, and an image composing step. The segmentation step and the segment-based alignment step can be formulated as an optimization problem. We design each cost function with smoothness term between frames. Experimental comparisons with synthesized and actual image sequences demonstrate that the proposed algorithm outperforms existing algorithms.
II. SHDR IMAGING

Here, we briefly review SHDR imaging [14]. Figure 1 shows the overview of the SHDR imaging. First, each LDR image captured under different exposure setting is segmented based on three intensity ranges: the saturated, the appropriate, and the blacked-out ranges. The threshold of each intensity range of each LDR image is designed so that the same-shaped regions appear in different exposure LDR images. Then, appropriate intensity range regions are aligned based on their shape, similarly to a jigsaw puzzle. Detailed algorithms of the segmentation and the shape-based alignment are described in [14]. After alignment, the SHDR image is composed using a Poisson image reconstruction [15].

III. PROPOSED SHDR VIDEO

We can generate the SHDR video by applying of the SHDR imaging algorithm [14] frame-by-frame. However, this naive extension of the existing SHDR imaging algorithm produces jitter artifacts, because the naive extension does not incorporate temporal relations among frames. Here, we propose a SHDR video generation algorithm considering temporal smoothness. Figure 2 presents an overview of the proposed SHDR video generation algorithm. The proposed SHDR video generation algorithm consists of a Graph-Cut segmentation of a data cube, disparity estimation, and image composing. Each exposure image sequence can be regarded as a data cube with which axes are x, y, and frame. Those data cube of exposure image sequences are segmented to three intensity ranges: blacked-out, appropriate, and saturated intensity ranges. The appropriate regions of different exposure image sequences are aligned based on the disparity estimation. The SHDR video is reconstructed by composing the aligned appropriate regions. We use Poisson image reconstruction [15] for image composition as well as the SHDR imaging. In the following subsections, we describe Graph-Cut segmentation of the data cube and disparity estimation.

A. Segmentation

We use a Graph-Cut algorithm to segment the data cube of each exposure image sequence. Figure 3 schematically shows the nodes and the edges of the cost function of the Graph-Cut algorithm for the segmentation of the data cube. The black dots represent the pixels or the nodes. The blue and red lines respectively represent spatial and temporal edges, which represent smoothness term. The mathematical expression of the cost function of the segmentation can be expressed as

$$J(L^m) = \sum_{f=1}^{F} \sum_{l \in V} J_d(l^f_v; x^f_v) + \alpha \sum_{f=1}^{F} \sum_{(u,v) \in A} J_s(l^f_v, l^f_u) + \beta \sum_{f=2}^{F} \sum_{l \in V} J_f(l^f_v, l^{f-1}_v) + \beta \sum_{f=1}^{F-1} \sum_{l \in V} J_f(l^f_v, l^{f+1}_v),$$

where $L^m$ stands for a set of labels of $m$-th exposure image sequence, $l^f_v$ represents a label of $v$-th pixel of $f$-th frame of $m$-th exposure image sequence, $x^f_v$ represents the scene radiance of the $v$-th pixel of the $f$-th frame of the $m$-th exposure image sequence, $F$ denotes the number of frames, $V$ represents a set of pixel numbers, $A$ represents a set of adjacent pixel pair, $\alpha$ and $\beta$ are tuning parameters, $J_d$ represents a data term, $J_s$ represents a spatial smoothness term, and $J_f$ represents a temporal smoothness term, respectively. Labels 0, 1, and 2 respectively represent the blacked-out, the appropriate, and the saturated.
The data terms for the saturate, the appropriate, and the blacked-out labels are shown in Fig. 4. Thresholds $\theta_1$ and $\theta_2$ are designed so that same-shaped regions appear in different exposure images. This relation between the threshold and the shape of region is the key of the segment-based disparity estimation. Detailed settings of thresholds are described in [14].

Spatial and temporal smoothness terms can be expressed as simple differences of labels, as $J_s(l_i^f, l_j^f) = |l_i^f - l_j^f|$, $J_f(l_i^f, l_j^{f-1}) = |l_i^f - l_j^{f-1}|$. These smoothness terms also contribute to noise reduction. The temporal smoothness is more effective than the spatial smoothness because the same location of adjacent frames tend to be same label if the motion of the target object is small. We empirically set the parameters as $\alpha = 2$ and $\beta = 4$.

B. Segment-based disparity estimation

Different exposure image sequences are taken by different viewpoints of cameras. We must align the appropriate regions of different exposure image sequences to reconstruct the SHDR video. For the alignment, disparities between the appropriate regions must be estimated. However, textures of some regions in the scene, which includes extremely bright and dark regions, are lost completely. Therefore, we estimate disparities based on the shapes of regions as jigsaw puzzle [14].

In segment-based disparity estimation, first, the image sequence that has the largest area of the appropriate regions is selected as a reference image sequence $I_{ref}^f$. The merged region of the appropriate and the saturated regions in the shorter exposure image $I_{ref-1}^f$ become the same shape as the saturated region in the reference image $I_{ref}^f$. As discussed in the previous section, the thresholds are designed to satisfy this relation. Therefore, we can estimate the disparity between the saturated region in the reference image and the merged region of the appropriate and the saturated regions in the shorter exposure image based on the region shape instead of the regions texture. It is a strong advantage for a scene which requires an extremely high-dynamic range. Once the disparity is obtained, the alignment of the regions is an easy task. Blacked-out regions in the reference image are also aligned in the same manner. This alignment algorithm is applied iteratively to align all appropriate regions in different exposure images.

The segment based disparity estimation can be formulated as an optimization problem. The cost function of this optimization problem is

$$E(T) = \sum_{j=2}^{F-1} \sum_{k=1}^{N_f} \left( \lambda_s E_s(l_j^f) + \lambda_f E_f(\Omega_j^f, t_j^f, t_{j-1}^f) \right),$$

where $N_f$ stands for number of segments in the $f$-th frame, $\Omega_k^f$ represents the $k$-th segment of the $f$-th frame, $\Omega_k^f$ represents the area of segment $\Omega_k^f$, $t_k^f$ denotes the disparity vector of segment $\Omega_k^f$. $T$ represents the set of the disparity vector $\{t_k^f : 1 \leq k \leq N_f \}$. $E_s(l_j^f)$ represents the shape similarity cost of the segment $\Omega_j^f$, $E_s$ denotes the smoothness cost of the spatially neighboring regions, $E_f$ stands for the temporal smoothness cost, and $\lambda_s$ and $\lambda_f$ are tuning parameters. Details of the shape similarity cost are presented in [14]. The smoothness cost of the spatially neighbor regions, $E_s$, can be expressed as

$$E_s(t) = \sum_{i,j \in B_f} w_d(\Omega_i^f, \Omega_j^f) ||t_i^f - t_j^f||^2,$$

where $B_f$ represents the set of neighbor regions in $f$-th frame, $w_d$ stands for a Gaussian weight based on the distance between the center of the gravities of two regions. The temporal smoothness cost, $E_f$, can be expressed as

$$E_f(\Omega_j^f, t_j^f, t_{j-1}^f) = \sum_{k=1}^{N_f} ||t_k^f - P_k(t_{j-1}^f, \Omega_k^f)||^2,$$
where \( P_k \) is the disparity prediction from the adjacent frame. An associated segment region is also moved and deformed if an object moves. Therefore, the disparity comparisons of the same segment region between different frames are not straightforward. In the proposed algorithm, the disparity of the segmented region in the current frame is predicted with the disparity map of the adjacent frame, as shown in Fig. 5. The segment mask in the current frame is overlaid on the disparity map of the adjacent frame. Then, the disparity is predicted by averaging the masked region of the disparity map. The disparity predicted by this manner is used to evaluate the temporal smoothness cost. Figure 6 presents a schematic diagram of the disparity estimation considering the temporal smoothness. First, the disparities of each frame are estimated in a frame-by-frame manner. Then, the disparity of the current frame is refined with the disparities predicted from the previous and the subsequent frame of disparity maps. This refinement is applied iteratively. For the example illustrated in Fig. 6, the car moves from the right to the left. The viewpoint difference causes a vertical disparity. The disparity depends on the depth of the object (car). The disparity of the object is unchanged even if the object moves with a constant depth. For this reason, we can assign a large weight for the temporal smoothness cost even for a dynamic scene, which makes the disparity estimation very stable.

C. Image composition

Once the disparity of each appropriate segment region is obtained, we can simply align the appropriate segment region. Some artifacts are present on the boundaries of the aligned segment. We compose the HDR image video using Poisson image reconstruction to reduce the artifacts on the segment boundaries [15].

IV. EXPERIMENTAL RESULTS

For evaluation, we show results generated using the proposed algorithm and compare the results with those of state-of-the-art algorithms. Video sequences are available at our project page\(^1\). First, we create a 3D computer graphics (CG) simulation dataset\(^2\) for more precise quantitative assessment to obtain the ground truth using Blender, 3D CG software. Blender can generate HDR videos and set camera positions and parameters to simulate videos taken in real scenes (Fig. 7). We produce LDR images at exposures of each camera. We assume that input videos are taken at six positions by Camera \( n (1 \leq n \leq 6) \). These camera positions are presented in Fig. 9. The distance between two adjacent cameras aligned vertically or horizontally is 30 mm. The exposure ratio between Camera \( n \) and Camera \( n - 1 \) is set as four times. Figure 7 shows a sequence of input images that is one of the input video frames. In Fig. 8, we present a comparison of the resultant two consecutive frames of our algorithm and state-of-the-art algorithms; Sen’s algorithm [1] and the SHDR imaging algorithm [14] applied frame-by-frame. In the result of Sen’s algorithm [1] saturated regions still remain. Although the SHDR imaging [14] produces a fine HDR image without saturated regions, it suffers from jitter artifacts observable in the video. On the contrast, the proposed SHDR video generation algorithm can yield temporally smoothed SHDR video.

For quantitative evaluation, we follow an SSIM-based approach proposed in [16]. The original SSIM [17] has three terms: the contrast, the structure, and the luminance terms. The range of the reconstructed HDR image strongly depends on the reconstruction algorithm. In other words, the reconstructed HDR has a scale ambiguity. In order to handle this ambiguity for the quality assessment, Ma et al [16] proposed to evaluate the quality of the HDR image with the contrast and the structure terms without the luminance term because the luminance is not regarded as a highly relevant component for the evaluation of the HDR image. We evaluate this two-term SSIM each frame. Figure 10 shows the two-term SSIM of Sen’s algorithm [1], the SHDR imaging algorithm [14], and the proposed algorithm. This results show the proposed algorithm outperforms the existing algorithms except first few frames.

Then, we take multi-exposure LDR videos simultaneously using an actual multiple camera system. The camera positions and exposures are the same as simulation setting (Fig. 9). We set exposures by changing gains, F-number and/or attaching ND filters without changing exposure time to keep motion blur effect unchanged for each image. Figure 11 presents the sequence of input images that is one of the input video

---

\(^1\)http://www.ok.ctrl.titech.ac.jp/res/SHDR/

\(^2\)We use 3D model from http://www.blendswap.com/blends/view/26649 and HDR background from http://noemotionhdrs.net/
frames. In Fig. 12, we present a comparison of the resultant two consecutive frames of our algorithm with the results of state-of-the-art algorithms; Sen’s algorithm [1], Heo’s algorithm [11], and the SHDR imaging [14] applied frame-by-frame. The results of Sen’s algorithm [1] and Heo’s algorithm [11] remains over- and under-saturated. The SHDR imaging misaligned the LED light behind the car. The proposed SHDR video generation algorithm can correctly generate the SHDR video. For quantitative evaluation, we use SSIM for multi-exposure fusion (MEF) [16], because the ground truth HDR image cannot be obtained. The SSIM-MEF is the quality assessment with a tone mapped HDR image and LDR images with different exposures. We can quantitatively evaluate the quality of the tone mapped HDR image without the ground truth of the HDR image. Figure 13 shows the evaluated SSIM-MEF for four algorithms; Sen’s algorithm [1], Heo’s algorithm [11], the SHDR imaging [14], and the proposed algorithm. This quantitative comparison demonstrates that the proposed algorithm outperforms existing algorithms.

V. CONCLUSION

We have proposed the super high dynamic range (SHDR) video generation algorithm. The proposed algorithm consists of region segmentation, disparity estimation, and image composition. We have introduced a temporal smoothness term to the region segmentation and the disparity estimation. These temporal smoothness terms improve the temporal stability and the precision of disparity estimation. Quantitative and qualita-
Fig. 11. Example frames of the real image sequence captured by six-camera system.

Fig. 12. Example frames of the reconstructed HDR video with the real image sequence.

Fig. 13. Evaluation of the SSIM-MEF with real image sequence.

tive evaluations conducted with a synthetic image sequence and a real image sequence demonstrate that the proposed algorithm outperforms existing algorithms.

REFERENCES


