J. Vis. Commun. Image R. 33 (2015) 309–322

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Depth video spatial and temporal correlation enhancement algorithm based on just noticeable rendering distortion model ☆

Zongju Peng [a,*], Fen Chen [a], Gangyi Jiang [a], Mei Yu [a], Feng Shao [a], Yo-Song Ho [b]

[a] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China
[b] School of Information and Mechatronics, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Republic of Korea

ARTICLE INFO

Article history:
Received 15 March 2015
Accepted 8 October 2015
Available online 22 October 2015

Keywords:
Three dimensional video system
Depth video
Just noticeable rendering distortion
Virtual view rendering
Spatial and temporal enhancement
Human visual perception
Depth video segmentation
Bit rate saving

ABSTRACT

Spatial and temporal inconsistency of depth video deteriorates encoding efficiency in three dimensional video systems. A depth video processing algorithm based on human perception is presented. Firstly, a just noticeable rendering distortion (JNRD) model is formulated by combining the analyses of the influence of depth distortion on virtual view rendering with human visual perception characteristics. Then, depth video is processed based on the JNRD model from two aspects, spatial and temporal correlation enhancement. During the process of spatial correlation enhancement, depth video is segmented into edge, foreground, and background regions, and smoothened by Gaussian and mean filters. The operations of the temporal correlation enhancement include temporal–spatial transpose (TST), temporal smoothing filter and inverse TST. Finally, encoding and virtual view rendering experiments are conducted to evaluate the proposed algorithm. Experimental results show that the proposed algorithm can greatly reduce the bit rate while it maintains the quality of virtual view.

1. Introduction

With development in the areas of image and video capturing technologies, three-dimensional (3D) display technologies and network infrastructures, the traditional two-dimensional video cannot satisfy the increasing demands of users. 3D video (3DV) systems provide viewers with depth perception of the observed scene and interactive interface. With these characteristics, 3DV systems have a broad prospect in consumer electronic applications such as mobile phones, games, cinema, and television [1,2]. In 3DV system, 3D scenes can be represented by two or more videos captured by different viewpoints. However, this representation method only provides fixed views, and users cannot continuously switch among views. Multiview video plus depth (MVD) is a more effective scheme of 3D scene representation, and it can be utilized to realize 3DV systems [3]. In an MVD-based 3DV system, MVD signals are captured, compressed and transmitted to the client. The decoded videos, along with videos of virtual views rendered by depth image based rendering (DIBR) technique [4,5], are displayed by 3D devices. MVD signals include multiple color videos and associated depth videos of the same scene. Depth video is geometric information of a scene, and denotes the distance between the captured scene and camera. Acquisition and transmission of depth video are two key techniques of realizing an MVD based 3DV system.

Recently, depth video coding has become an active research focus. The most straightforward method is to compress depth video using most advanced encoding standard [6,7]. However, depth video exhibits the following particular characteristics compared with conventional color video: (1) depth video is monochrome, (2) the texture of depth video is not as rich as that of the corresponding color video and (3) depth video is not used for display, but rather for virtual view rendering. Many related technologies were proposed for depth video/map encoding [8–20]. Platelet-based [8], silhouette-based [9], enhanced context-based adaptive binary arithmetic [10], and object-based coding algorithms [11] were proposed for improving encoding performance. Liu et al. presented two depth compression techniques: trilateral filter and sparse dyadic mode [12]. Kang et al. designed an adaptive geometry-based intra prediction scheme for depth video coding [13]. Oh et al. proposed a depth boundary reconstruction filter to code depth video [14]. Spatial down sampling and temporal subsampling approaches are utilized to improve compression ratio [15,16]. Bosc et al. investigated factors which impact on the best bit rate ratio between depth and color video [17]. Zhang et al. proposed regional bit allocation and rate distortion optimization

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author. Fax: +86 057487600582.
E-mail address: pengzongju@nbu.edu.cn (Z. Peng).

http://dx.doi.org/10.1016/j.jvcir.2015.10.003
1047-3203/

algorithms for multiview depth video coding by imbalance bit rate allocation for different regions [18]. Zhang and Shao respectively proposed depth video coding algorithms based on distortion analyses [19,20].

Accurate and consistent depth video is the foundation of high compression performance of depth video coding. In general, depth acquisition methods include depth extraction from computer graphic content [21], depth from structured light [22], Kinect sensors [23], depth camera system [24,25] and depth estimation software [26]. Limited by the principle of Kinect sensors based on structured light technique, depth images suffer from temporal flickering, noise, holes and inconsistent edges between depth and color images [23]. In depth camera system, depth video is captured based on the principle of time-of-flight [24,25]. The captured depth maps may be inconsistent with the scene because of ambient light noise, motion artifacts, specular reflections, and so on. In addition, depth camera is too expensive to use on a large scale. So far, depth estimation software is the alternative method of depth map acquisition [26]. However, depth video obtained by depth estimation software usually contains discrete and rugged noises. Hence, depth video is inaccurate and inconsistent. The temporal and spatial correlation is weak so as to decrease the compression performance.

In order to improve encoding and rendering performance, many depth video processing algorithms [27–34] have been proposed. Mueller et al. produced accurate depth video for artifact-free virtual view synthesis by combining hybrid recursive matching with motion estimation, cross-bilateral post-processing and mutual depth map fusion [27]. Min et al. presented a weighted mode filtering method that enhances temporal consistency and addressed the flicking problem in virtual view [28]. Nguyen et al. suppressed coding artifacts over object boundaries using weighted mode filtering [29]. Ekmekcioglu et al. proposed a content adaptive enhancement method based on median filtering to enforce the coherence of depth maps across the spatial, temporal and inter-view dimensions [30]. Kim et al. presented a series of processing steps to solve the critical problems of depth video captured by depth camera [31]. One of these processing steps is the enhancement of temporal consistency by an algorithm based on motion estimation. Zhao et al. proposed a depth no-synthesis-error (D-NOSE) model and presented a related smoothing scheme for depth video coding [32]. Fu et al. proposed a temporal enhancement algorithm for depth video by utilizing adaptive temporal filtering [33]. In our previous work, depth video was enhanced by temporal pixel classification and smoothing [34].

However, depth video processing algorithms [27–34] do not consider the perception of human visual system (HVS), and still leave room for improvement. Zhao et al. proposed a binocular just-noticeable-difference model to measure the perceptible distortion of binocular vision for stereoscopic images [35]. Silva et al. experimentally derived a just noticeable depth difference (JNDD) model [36] and applied it to depth video preprocessing [37]. Jung proposed a modified JNDD model that considers size consistency, and then used it in depth sensation enhancement [38]. The JNDD models are built by subjective tests on stereoscopic displays. Hence, they are display dependent, and not suitable for estimating depth distortion range in virtual view rendering. In this study, we propose a just noticeable rendering distortion (JNRD) model and apply it for spatial and temporal correlation enhancement. Different from other distortion models [35,36,38], the JNRD model is built based on the perception distortion of virtual view. Firstly, the JNRD model is formulated by combining geometric displacement in DIBR with just noticeable distortion (JND) model that reflects human visual perception. Then, the spatial and temporal correlation of depth video is enhanced using the JNRD model. Finally, the proposed algorithm is appraised from three aspects, compression ratio, objective quality and subjective quality of the virtual view. The experimental results show that compression performance of the processed depth video is improved in comparison with the original depth video and the processed video of other algorithms. The proposed algorithm maintains quality of the virtual view.

The rest of this paper is organized as follows. Section 2 describes the problem of the depth in spatial and temporal correlation, and presents the overall block diagram of the proposed algorithm. Section 3 describes the JNRD model. Sections 4 and 5 present the spatial and temporal enhancement algorithm in detail. Experimental results are given in Section 6. Finally, conclusions are made in Section 7.

2. Proposed depth video correlation enhancement algorithm

2.1. Problem description

Depth video is inconsistent along spatial and temporal directions because of the limitations of depth video capture technologies. Fig. 1 shows the frames and frame difference of the color video and the corresponding depth video in 'Leave Laptop' sequence. The frame S_iT_j denotes the frame in the ith view at the jth time instant in the video sequence. Fig. 1(a), (b), (d), and (e) are the frames $S_{10}T_{35}$ and $S_{10}T_{36}$ of color video, and the associated depth frames, Fig. 1(c) and (f) are the texture and the depth frame difference images between frames $S_{10}T_{35}$ and $S_{10}T_{36}$ where black means larger difference. The scene in the red rectangular region of the color video is nearly at the same imaging plane. Correspondingly, the depth value in the corresponding region should be nearly the same. However, the depth value in the corresponding region is not consistent with the corresponding color video. Depth inconsistency decreases the spatial correlation of depth video.

Depth video inconsistency also decreases temporal correlation. In the scene in Fig. 1, only the men and chair are seen moving slightly. Hence, nearly total frame difference image of the color video, with the exception of the men and chair in the scene, is dark, which represents the content consistency along the temporal direction. In contrast, some areas in the frame difference image of the depth video, e.g., the blue rectangular region in Fig. 1(f), are dark, which means temporal inconsistency.

Consequently, depth video inconsistency eventually deteriorates encoding performance because the spatial and temporal correlation is the theoretical basis of high compression efficiency of video signals.

2.2. Proposed depth video correlation enhancement algorithm

To improve the compression performance of depth video, a new spatial and temporal correlation enhancement algorithm is proposed in this paper. Fig. 2 shows the block diagram of the proposed algorithm, which includes three parts, JNRD model building, depth video spatial correlation enhancement, and depth video temporal correlation enhancement. The JNRD model is the basis for depth video spatial and temporal correlation enhancement. In Fig. 2, $\mathbf{G}$ is the JNRD of the corresponding depth video; $\mathbf{R}$, $\mathbf{D}$ and $\mathbf{D}''$ are color video, original depth video and processed depth video, respectively; $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{B}$ are the edge, foreground and background regions of the depth video, respectively; and $\mathbf{D}'$, $\mathbf{D_1}'$, and $\mathbf{D_2}'$ are intermediate processing results of depth video.

In the proposed algorithm, the JNRD model of depth video is built firstly. Then, the depth video is processed in the order of spatial and temporal correlation enhancement. JNRD model building, spatial and temporal correlation enhancements are detailed in Sections 3, 4 and 5, respectively.