J. Vis. Commun. Image R. 38 (2016) 351–366

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Virtual view synthesis using layered depth image generation and depth-based inpainting for filling disocclusions and translucent disocclusions ☆

Suryanarayana M. Muddala, Mårten Sjöström *, Roger Olsson

Dept. of Information and Communication Systems, Mid Sweden University, 85170 Sundsvall, Sweden

ARTICLE INFO

Article history:
Received 16 April 2015
Accepted 23 February 2016
Available online 7 March 2016

Keywords:
View synthesis
Depth-image based rendering
Image inpainting
Texture synthesis
Hole filling
Disocclusions
Translucent disocclusions
Layered depth image

ABSTRACT

View synthesis is an efficient solution to produce content for 3DTV and FTV. However, proper handling of the disocclusions is a major challenge in the view synthesis. Inpainting methods offer solutions for handling disocclusions, though limitations in foreground-background classification causes the holes to be filled with inconsistent textures. Moreover, the state-of-the art methods fail to identify and fill disocclusions in intermediate distances between foreground and background through which background may be visible in the virtual view (translucent disocclusions). Aiming at improved rendering quality, we introduce a layered depth image (LDI) in the original camera view, in which we identify and fill occluded background so that when the LDI data is rendered to a virtual view, no disocclusions appear but views with consistent data are produced also handling translucent disocclusions. Moreover, the proposed foreground-background classification and inpainting fills the disocclusions with neighboring background texture consistently. Based on the objective and subjective evaluations, the proposed method outperforms the state-of-the art methods at the disocclusions.

1. Introduction

Three Dimensional Video (3DV) technologies offer an immersive user experience. In principle, 3DV in the stereo format includes two videos of the same scene but from slightly different viewpoints. The two views are then presented to the left and right eyes through separate channels. With the development of 3D display technology, auto stereoscopic displays are now available in the market [1]. In contrast to the stereo displays and current cinema technology, auto stereoscopic displays create depth impression without any additional eye wear. Free viewpoint television (FTV) is among the emerging applications of 3D video, which has even surpassed popularity of 3DTV in the topic of 3D video and content generation [2]. 3DTV provides the depth-impression using stereo, whereas FTV offers the viewer to look around the scene. In order to provide such an experience, these technologies require a number of camera views captured from different viewpoints. Usually the capturing and transmission of a big number of views is not a feasible solution, and so view synthesis is employed as an alternative solution.

Technically, virtual views are generated by using the principle of perspective geometry. An illustration of virtual view generation is shown in Fig. 1. In this respect, video-plus-depth (V + D) and multiview video-plus-depth (MVD) are common 3DV formats to efficiently transfer 3D video to the end user [3,4]. Later, additional views can be produced at the end user side by considering the requirement of the display. Depth-Image-Based Rendering (DIBR) is a widely used method to render a new viewpoint, using a texture and depth information [5]. We define the inputs to the DIBR as the original views and outputs as the virtual views (aka rendered views). The virtual view consists of the warped texture and depth images. True depth at virtual view is the depth obtained by some capture method.

As a consequence of DIBR, the warped images exhibit artifacts, namely *ghosting* and *uncovered areas* (aka *holes*), which affect the visual quality severely (see Fig. 2(a)). Ghosting artifacts are a mixture of colors at the edges that are projected into the neighboring objects due to the depth and texture misalignment at the depth discontinuities. *Holes* can be classified into *cracks*, *disocclusions* and *out-of-field areas*. Cracks are usually 1 to 2 pixel-wide missing information, which appear due to rounding the projected pixel position to the nearest integer. In the case when cracks appear on an object and that object causes an occlusion, texture from occluded objects seeps through the cracks and so we call them

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail address: marten.sjostrom@miun.se (M. Sjöström).

http://dx.doi.org/10.1016/j.jvcir.2016.02.017
1047-3203/

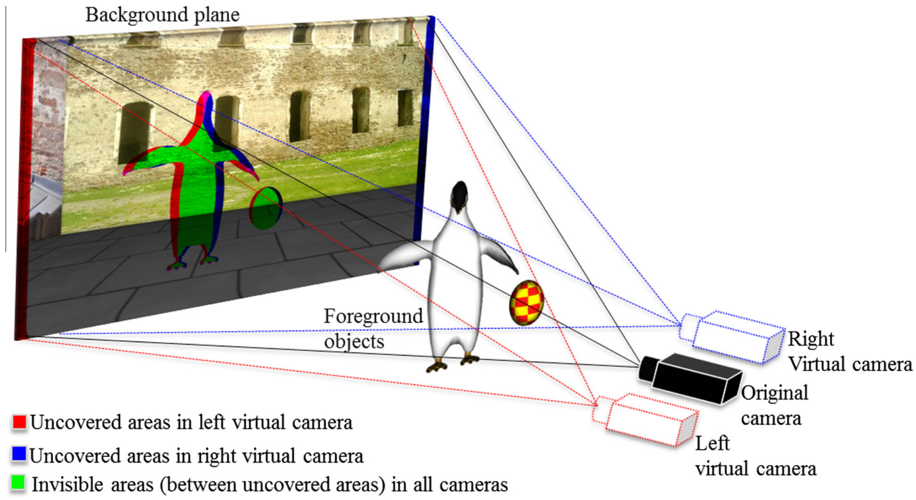

Fig. 1. Illustration of view-extrapolation of virtual cameras from original camera. Uncovered areas are invisible in original camera but visible in the virtual cameras, so they should be filled.

Fig. 2. View synthesis results: (a) synthesized image with holes (in yellow color) and (b) proposed synthesized image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

translucent cracks. Image processing techniques with different complexity levels have been developed to handle these artifacts, yet handling disocclusions and out-of-field areas are still challenging. Disocclusions are the result of baseline (distance between the cameras) and a distinct change in the depth between neighboring pixels, normally occurring at object borders. In this context, we define foreground (FG) to be the part of the scene closer to the camera that occludes other objects and background (BG) to be the part of the scene farther from the camera which is partially occluded by foreground in the original image. Out-of-field areas are caused in the virtual camera views due to the lack of information at the image boundaries. Generally holes get larger when the baseline increases. But large holes are not only involved in a large-baseline setup like FTV. They also appear when there is a large depth discontinuity between FG and BG, which increases the importance of efficient handling of the holes.

Besides the above mentioned artifacts, there exists another type of artifact, which we define as the *translucent disocclusion*. Translucent disocclusions differ from common disocclusions by exposing texture information that is present behind the BG. Translucent disocclusions are only visible when the depth has three or more layers, and where there are occlusions present between layers. An example of this case is shown in Fig. 3(a) in the area near the woman's hand. Note in Fig. 3(c) that the wall texture is seeping through the woman's body. The magnitude of the disturbance created by translucent disocclusion artifacts depends on the placement of occlusions in the scene and the occlusion area. These artifacts should be addressed especially if they have a large size, since that severely affects the perceived visual quality. Conventional *hole-filling* methods aim at the disocclusion artifacts and leave translucent disocclusion unattended, which consequently degrades the virtual view quality (see Fig. 4(b)–(f)).

1.1. RelatedWork

Several methods have been proposed in the literature to reduce the artifacts in the virtual view. Ghosting artifacts are reduced by first detecting the depth discontinuities by estimating the FG and BG contribution and removing pixels in the vicinity of the discontinuities [11,12]. We can classify the hole-filling methods into two categories: *Pre-processing* and *Post-processing*.

Pre-processing: Cracks are commonly reduced by using backward warping approach [13]. In this approach, first the depth