•Review•

A review on image-based rendering

Yuan CHANG^{1,2}, Guo-Ping WANG^{1,2*}

1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;

2. Beijing Engineering Research Center for Virtual Simulation and Visualization, Beijing 100871, China

* Corresponding author, wgp@pku.edu.cn

Received: 15 September 2018 Accepted: 13 November 2018

Supported by National Natural Science Foundation of China (61632003).

Citation: Yuan CHANG, Guo-Ping WANG. A review on image-based rendering. Virtual Reality & Intelligent Hardware, 2019, 1(1): 39—54
DOI: 10.3724/SP.J.2096-5796.2018.0004

Abstract Image-based rendering is important both in the field of computer graphics and computer vision, and it is also widely used in virtual reality technology. For more than two decades, people have done a lot of work on the research of image-based rendering, and these methods can be divided into two categories according to whether the geometric information of the scene is utilized. According to this classification, we introduce some classical methods and representative methods proposed in recent years. We also compare and analyze the basic principles, advantages and disadvantages of different methods. Finally, some suggestions are given for research directions on image-based rendering techniques in the future.

Keywords Image-based rendering; Virtual reality; Image interpolation; Panorama

1 Introduction

Virtual reality technology uses computer to construct a virtual environment in three-dimensional space, providing users with a high-fidelity, immersive visual, auditory and even tactile experience through display devices such as screens and helmets. Users can use interactive devices to interact with objects in the virtual world to achieve an immersive feel. Nowadays, with the rapid development of virtual reality technology, virtual reality has been widely used in many fields such as industry and entertainment.

The modeling and rendering of virtual scenes, namely the construction and representation of virtual objects in 3D virtual environment and the rendering of free viewpoint images during user roaming, are fundamental problems of virtual reality technology. Depending on the application scenario, the virtual environment can be a completely fictional three-dimensional scene, or it can be a reproduction of a real-world scene, or a three-dimensional environment created by adding fictitious elements based on real-world scenes. With the development of the internet, sensors, drones, etc., people can collect image data or video data from a large number of real-world scenes more easily. Therefore, how to use such data to reconstruct the information of the real scene in the virtual environment, so that users can roam the virtual scene from any viewpoint, has become a concern in virtual reality technology.

The traditional way to simulate real scenes is 3D modeling and rendering. The modeling part is to recover the geometric information of the scene and the illumination information. Common modeling methods include manual modeling using geometric modeling software, reconstruction of the geometry from multiple photographs capturing real-world scenes using structure-from-motion and texture mapping^[1,2]. The rendering part including geometric cropping, illumination calculation and element

www.vr-ih.com

simplification of the reconstructed scene according to the user's viewpoint, is to generate the image observed from the current viewpoint. This traditional process based on 3D modeling and rendering can fully express the geometric information of the scene, so that users can flexibly change the external conditions such as illumination to obtain different rendering effects, and can easily roam the scene freely. However, the complexity of this modeling and rendering process is very high. The time required for modeling and rendering will increase significantly, when the scale of the scene is larger and the surface details of the objects in the scene are richer. In addition, the rendering results obtained by this traditional 3D modeling and rendering process are usually not photo-realistic without high computational cost.

Due to these limitations of traditional 3D modeling and rendering methods, people are more concerned about whether they can roam the scenes directly using the collected image dataset. Google Street View captures street scenes using street view cars, and provides 360° panoramic photos for users so that they can see the street scenes from different directions^[3]. However, users can not look at the scene from arbitrary viewpoints because the image seen from uncaptured viewpoints can not be rendered, which is much easier using traditional 3D modeling and rendering process. Therefore, how to use finite and discrete captured image to obtain continuous and arbitrary viewpoints becomes an important issue. This is called image-based rendering. More specifically, image-based rendering technique is a technique of generating a rendering result of an unknown viewpoint by interpolating through discrete input images or re-projecting pixels in input images. This rendering method uses images as the fundamental elements, which requires no more complicated geometric modeling processes and rendering processes but only some analysis and processing of images. The complexity does not increase with the complexity of object relationships and surface details in the scene, which can significantly save computing resources. Moreover, since the results are directly rendered from photos, it can be photorealistic.

The essence of image-based rendering technology is to obtain all the visual information of the scene directly through images. Previous image-based rendering techniques can be classified into three categories according to how much geometric information is used: rendering without geometry, rendering with implicit geometry, and rendering with explicit geometry^[4]. Likewise, we classify previous techniques into two categories, rendering with or without geometry, to introduce some classic approaches and some new algorithms. We also analyze the limitations and key issues of these approaches.

2 Rendering without geometry

Rendering methods which do not utilize 3D geometric information are usually based on the synthesis of panoramas or panoramic videos, often requiring the cameras to have certain layouts. In order to introduce clearly the rendering methods which do not require 3D modeling, we first introduce plenoptic function, which is the theoretical basis based on image rendering. Plenoptic function is a function used to describe the intensity of each light ray in the scene. Image-based rendering technique tries to find some discrete samplings of the plenoptic function. Approaches which do not utilize the three-dimensional geometric information of the scene tend to simplify the form of the plenoptic function by limiting the pose of the camera, which makes it simple and feasible to recover the continuous form of the plenoptic function by discrete sampling of it. We first introduce the concept of plenoptic function and its significance in the rendering process. Then through some specific image-based rendering algorithms, we introduce different simplified forms of the plenoptic function and their advantages and disadvantages.

2.1 Plenoptic function

Rendering is the process of describing the scene information observed at a certain viewpoint. Traditional

rendering process expresses the observed scene information through the geometric model of the scene, the light illumination model, the light source, the texture and other information. No matter the human eye or the camera, the capturing of scene information is done by receiving light rays emitted, reflected or refracted by objects in the scene. From this perspective, the traditional modeling and rendering process generates the image captured by the human eye or the camera by simulating the light source in the scene, the position of the object, and the nature of the reflection. Moreover, directly simulating the human eye or the camera is also a feasible way of simulating the scene information observed at the observation point. The idea of image-based rendering is actually to simulate the receiving end directly. The mathematical expression of this kind of simulation is plenoptic function.

The plenoptic function describes the intensity of light passing through any certain observation point in any certain direction. Complete plenoptic function is a 7-dimensional function^[5,6], defined with observing location (V_x, V_y, V_z) , angle of incidence (θ, ϕ) , the wavelength λ and the time t . When considering static environment with fixed lighting condition, we can drop out two variables, time and wavelength. Then we get a 5-dimensional plenoptic function^[7]:

$$P_5 = P(V_x, V_y, V_z, \theta, \phi)$$

If V_x, V_y, V_z are fixed, the plenoptic function describes a panorama at fixed viewpoint (V_x, V_y, V_z) . A regular image with a limited field of view can be regarded as an incomplete plenoptic sample at a fixed viewpoint.

With plenoptic function, we can redefine the image-based rendering problem, that is estimating more complete plenoptic function using discrete sampling. How to sample and how to construct the plenoptic function are both important issues. The more constraints on the camera position, the simpler the form of the plenoptic function becomes. For example, as long as we stay outside the convex hull of an object or a scene, if we fix the location of the camera on a plane, we can use two parallel planes (u, v) and (s, t) to simplify the complete 5D plenoptic function to a 4D lightfield plenoptic function^[8,9] (Figure 1):

$$P_4 = P(u, v, s, t)$$

Figure 1 Plenoptic function represented by 4D light fields^[8].

2.2 Cylindrical panoramic image mosaic

Panoramic image mosaic is a common approach to help free viewpoint navigation. A plenoptic function describes a panorama when the viewpoint is fixed. We can render images captured from any directions by constructing panoramas^[10,11]. Here, taking the cylindrical panorama as an example, we introduce image-based rendering approaches using panoramas.

With cylindrical panorama, we can freely change our angle of view horizontally. Panoramic photos can be captured with specialized cameras^[12], and they can be obtained by image stitching as well. As we mentioned above, a plenoptic function describes a panorama when the viewpoint is fixed. Therefore, we can get a complete plenoptic function by constructing a panorama by image stitching. If the camera's intrinsic matrix is fixed, we can project each regular images to a cylindrical surface to obtain the cylindrical panorama (Figure 2). Shown in Figure 2, given the coordinate (x, y) in a regular image, we can directly obtain the corresponding (u, v) parameters in the cylindrical environment map. In practice, there may be distortion because of the accumulated registration errors. To solve this problem, people combine both global and local alignment, which significantly improves the quality of the image mosaics^[13], to obtain

a better panorama (Figure 3).

Cylindrical panorama can only help roaming where the viewpoint is fixed at a certain position. Users can look at any directions, but they can't roam the scene by moving forward and backward, etc. Google Street View makes a simple transformation between panoramas of adjacent sampling locations by densely collecting panoramic views of street scenes, allowing users to convert from one viewpoint panorama to another viewpoint panorama. But the conversion process is very rough and can't provide users with a realistic experience of roaming the scene^[3].

Figure 2 Projection from a regular image to a cylindrical surface.

Figure 3 A panoramic image stitched from some photographs^[11].

2.3 Image mosaic through camera rotation

Concentric Mosaics is a 3D parameterization of the plenoptic function^[14], where the camera motion is constrained along concentric circles on a plane. Concentric mosaics can be created by compositing slit images taken at different locations of each circle. Here, a slit image is a single vertical-line image, which can be simulated with a regular camera by keeping only the central column of pixels of each frame. There is a possible concentric mosaic capture system setup, where a number of cameras are mounted on a rotating horizontal beam that is supported by a tripod. Each camera is constrained to move continuously and uniformly along its circle. We take the central column of pixels of image captured by each camera as the slit image, and we put them together to obtain the concentric mosaic. By moving the camera further away from the rotation axis, we obtain a collection of concentric mosaics (Figure 4). For each circle, we would like to capture both viewing directions at each tangent line. This can be done, for example, with two cameras facing the opposite directions.

Concentric Mosaics simplify the plenoptic function to three parameters, radius, rotation angle and vertical elevation. Different from cylindrical panorama, it allows the viewer to explore a circular region and experience horizontal parallax and lighting effects. Novel views are created by composing slit images

in the direction of the desired frustum. As the radial distance between two adjacent circles increases, the number of slit images available for reconstruction of a given novel view is reduced, requiring more interpolation of the original data to fill in the image plane of the novel view.

Concentric Mosaics have much smaller file size because only a 3D plenoptic function is constructed. In addition, Concentric Mosaics are very easy to capture. However, because off-the-plane rays can not be synthesized correctly without knowing corrected depth information, there will be vertical distortions in the rendered images. Although vertical distortions can be reduced with constant depth correction, they still exist.

Unlike Concentric Mosaics, Debevec et al. construct a spherical light field capture system to help free viewpoint navigation^[15]. It records outward-looking spherical light fields using fisheye camera on a programmable pan/tilt head. The system shoots 72 images horizontally by 24 images vertically, spaced 5° centered around the equator, resulting in a set of 1728 images (Figure 5). Since the fisheye covers a hemisphere of 180° , this system provides the data necessary to render views of the scene from any position within the sphere outlined by the cameras. When rendering using HMD, the images are resized to the resolution appropriate for device first, and then render two views for both eyes individually.

Figure 4 A collection of concentric mosaics^[14].

Figure 5 Spherical light field capture system^[15].

This approach can provide free-view roaming within a certain range, but viewpoints that are too far from the center of rotation cannot be rendered. What is more, the capturing process is complicated, and the cost of data acquisition is relatively high.

3 Rendering with geometry

Due to the diversity and complexity of real-world scenes, when image-based rendering is not performed using 3D geometric information of the scene, the spatial relationship of the objects in the scene cannot be completely described, so the rendering result has various restrictions, such as the camera's spatial position is limited, the image collection density can not be small, and the occlusion in the scene cannot be handled. In recent years, with the development of three-dimensional information acquisition of scenes from images, 3D information contained in the images can be extracted and described in many ways, which also promotes the development of image-based rendering techniques. People began to care about how to use the three-dimensional information contained in the image to help image-based rendering, so as to solve various problems that may occur only using color information from images, and obtain higher quality free viewpoint rendering results.

It should be noted that, unlike the traditional 3D modeling and rendering process, in image-based rendering process, there is no need to perform texture mapping and rendering for 3D models, or even no need to get the dense 3D point cloud. Image is still the basic element of the rendering process, and the 3D structure is only used as an auxiliary information to help interpolate the image of unknown viewpoint directly from the image pixels. Different methods of image-based rendering use different forms of three-dimensional information. For example, they can be 3D models represented by triangle mesh^[16–18], dense/sparse point cloud reconstructed from multiple images^[19,20], depth maps or depth distributions^[21–23], or even only correspondences^[24–26].

Ideally, when the precise three-dimensional information of the scene is obtained, the color of each pixel of the input image can be easily projected according to the three-dimensional information to a certain fixed position in 3D space. Then according to the camera parameters of the novel viewpoint, the color of the space point can be projected onto the image plane of the novel view, so that the new view can be rendered. In practice, however, this often fails to get the right rendering results, because (1) not all pixels of all the input images can find the corresponding spatial positions; (2) even if each pixel finds a corresponding spatial location, the spatial locations corresponding to different pixels (in the same image or different images) representing the same object may be inconsistent. Therefore, after obtaining the representation of the three-dimensional information of the scene, further processing is needed to solve the above problems. Many image-based rendering methods are proposed to solve the above two problems. Some algorithms propose methods for complementing imperfect three-dimensional information^[20,25,26], and some algorithms deal with inconsistencies in the three-dimensional information inside one image and between different images^[20,23].

The two problems mentioned above are mainly caused by the errors induced by the reconstruction of the three-dimensional structure of the scene. In addition, there are some difficulties because it is difficult to recover the three-dimensional information due to some special structures in the scene, such as reflection and thin structures (Figure 6). Some approaches have been proposed to solve these special cases^[27–29]. Next, some typical image-based rendering methods based on various key problems and using different three-dimensional information forms will be introduced and analyzed.

Figure 6 Thin structures are difficult to reconstruct^[29].

3.1 View interpolation based on image correspondence

We introduce in this section image-based rendering methods using image correspondence as geometry information. For two input images, Chen and Williams use dense optical flow between them to interpolate arbitrary viewpoints^[24]. This method can get better results for the case where the two input images are relatively close, because when the viewpoints are relatively close, the visibility of the objects in the scene has less influence on the result of interpolating the new viewpoint images. However, when the difference between the input images is large, the overlapping area of the two images is also reduced, and the result of interpolation by this method is deteriorated. Figure 7 shows that this method performs well for close images.

With the development of the dense correspondence estimation method, there are many solutions to the matching problem between some images with large differences in viewpoints^[30,31]. Moreover, even for image pairs that are difficult to directly find a complete dense correspondences, the interpolation problem

Figure 7 Images interpolated using correspondence (two middle ones are interpolated images)^[24].

between images with large difference can be handled by complementing the incomplete correspondences. Nie et al. proposed a method dealing with the problem that the correspondences between the image pairs with large differences is not complete enough to interpolate the new viewpoints^[26]. The approach first over segments the image to super-pixels^[32], ensuring that each super-pixel represents the same plane. Then it uses correspondence estimation method to calculate a homography for each super-pixel to characterize the motion relationship between two images of the super-pixel. Due to the incompleteness of the correspondences caused by the large difference of viewpoints, only a part of the super-pixel's homography can be fitted by the RANSAC algorithm through enough correspondences in the super-pixel^[33]. For super-pixels without sufficient matches or with poor matching consistency, homographys are obtained by propagation of other super-pixels similar and nearby.

This method of performing viewpoint interpolation based on image correspondences does not require complex three-dimensional reconstruction of the scene, nor does it depend on an accurate depth map. It is suitable for smooth transition of viewpoints from image to image, and has the advantage of low computational cost. However, since the depth relationship between objects in the scene is not considered, the occlusion problem cannot be handled. When there is a complex occlusion relationship in the scene, ambiguity may occur, which may affect the quality of the rendering result.

3.2 Rendering methods with incomplete depth map

Some approaches are based on the problems caused by incomplete image depth recovery. Due to the complexity of the scene and the reconstruction error, sometimes the depth value of only partial area of the image can be calculated. Restoring a complete depth map with these sparse depth value samples has always been a concern. Some methods reconstruct a dense depth map by performing high-density sampling of depth values on the contour regions of objects in the image^[34], while some methods obtain a dense depth map by performing a piecewise plane fitting method on the point cloud^[35–37]. Some of these depth synthesis methods are well suited for image-based rendering problems.

Chaurasia et al. have proposed an image-based rendering method^[19], in which a depth synthesis algorithm is proposed to provide depth sampling for areas with poor depth recovery in the image. A rendering algorithm based on local shape-preserving warping is also proposed to deal with the inconsistency caused by depth inaccuracy. The problem is to get a better visual result.

First we introduce how the algorithm gets depth value samples. In order to obtain the depth value sampling, the camera parameter of each image can be extracted using multiple view geometry methods from several input images^[1], and the 3D point cloud can be reconstructed^[2]. After the reconstruction result is obtained, the point cloud is re-projected back to each input image to obtain a depth value sample for each image. Since it is not guaranteed that each region in the image can be corresponding to a part of the 3D point cloud, only part of the pixels will be re-projected by points in the point cloud. So only some pixels of each image have depth values (Figure 8).

Figure 8 (a) Input image; (b) Projected depth samples^[19].

When completing the depth map, it first over-segments the image into super-pixels^[32]. Then using the hypothesis that pixels in the same super-pixel have depth consistency, it constructs a weighted undirected graph with super-pixels as nodes, and uses super-pixels containing sufficient depth value samples to synthesize the depth of super-pixels without sufficient samples. Therefore, a complete depth map is obtained. In addition, in order to deal with the rendering error caused by the error induced by the point cloud noise, the algorithm also performs a shape-preserving warping for each super-pixel. It finds the bounding box along the coordinate axis for each super-pixel, and constructs a triangle mesh in the bounding box. Then it uses the constraint of the mesh to warping the super-pixel, so as to obtain a better visual rendering result.

The method uses depth synthesis to propagate the depth of other super-pixels to the super-pixels without sufficient depth samples, which solves the problem of incomplete geometric information to some extent. However, this way of synthesizing the depths has an obvious drawback, that is, when the super-pixel missing depth information cannot find other super-pixels with sufficient depth samples belonging to the same object in the image, the wrong depth value will be synthesized. If an object in the scene is completely missed during the reconstruction, the correct rendering result of the object cannot be obtained. In addition, this method relies heavily on the results of super-pixel segmentation because the depth synthesis algorithm is based on the assumption that the depth values of pixels within the same super-pixel are consistent. So if there are small, thin objects in the scene, so that the super-pixel segmentation cannot split it, the rendering result will also be incorrect in this area.

3.3 Rendering based on geometry consistency

In the area of image-based rendering using three-dimensional information, in addition to the above-mentioned problem that the three-dimensional information recovery is not complete enough, the problem of inconsistency of three-dimensional information is important as well. It includes inconsistencies in depth of adjacent pixels within the same image or in depth between adjacent images due to reconstruction errors.

3.3.1 Rendering using dense correspondence and depth map

Lipski et al. proposed an image-based rendering method which combines the usage of dense correspondence and depth map^[20]. They apply image morphing in three-dimensional space, and compensate for inaccuracies and invalid assumptions made in reconstruction by aligning image regions according to bidirectional correspondence maps.

They proposed an iterative scheme for depth and correspondence estimation. First, using established SFM techniques, they estimate camera calibration and a sparse 3D point cloud representation of the

scene^[1,2]. For every pixel x in a reference image A, we query the 3D scene structure if a point exists that projects to it. If such a point exists, we can derive a correspondence vector estimate $\tilde{w}_{AB}(x)$ for any neighboring view B. We call $\tilde{w}_{AB}(x)$ geometry-guided correspondence. The geometric information can be exploited to constraint the correspondence search. The correspondence search is based upon an optimization scheme that minimizes an energy term. Different from traditional optimization scheme^[38,39], they add an additional geometric constraint that enforces consistency between the geometric information and the estimated correspondences. After obtaining the dense correspondence, they triangulate an actual world space position if they know at least one valid corresponding pixel in another image by intersecting both viewing rays. Correct depth value cannot be obtained for invalid pixels, so they segment each image into super-pixels and fit a plane for each super-pixel to help obtaining a robust result. After depth map is calculated, the geometric-guided correspondence can be updated, and the dense correspondence can be optimized again.

After obtaining a dense correspondence between images and the depth map of each image, the pixels in each image can be projected to three-dimensional points in the space. However different pixels representing the same point in the real scene may be not projected to the same 3D point due to the error of the depth value. But according to the correspondence map, it can be known that which 3D points correspond to the same 3D point in the real scene. Then the unified 3D coordinates can be calculated according to these points, and projected to the image plane of the novel view (Figure 9).

Figure 9 Correspondence and depth based rendering.

The method uses the dense correspondence between images as a condition to reduce the inconsistency of depth information between different images caused by reconstruction errors, which plays an important role in reducing the ghost in the rendering result. However, similar to other methods relying on super-pixel segmentation, the rendering quality of the novel viewpoint is greatly affected by the quality of the super-pixel segmentation. In addition, for input images with a large difference in viewpoint, the obtained depth and the quality of the correspondence are also degraded, thereby affecting the rendering effect. Longer calculation time is also a disadvantage of this method. Because of the need to densely match two adjacent images, when the input images are too many and the image resolution is high, it may take a lot of time to get the correspondence by optimizing the energy. And because the depth map and correspondence are iteratively optimized, the correspondence estimation between every two images is calculated more than once, which greatly increases the time consumption.

3.3.2 Soft 3D reconstruction for image-based rendering

Penner et al. proposed another image-based rendering method^[23], which considers the inconsistency as well, that is, performing soft reconstruction of the scene. Unlike depth map which gives a fixed depth value to each pixel, or point cloud which reconstructs a fixed position by multi-view geometry method, soft 3D reconstruction constructs a distribution of depth values for each pixel. Based on this expression of the three-dimensional scene, a similar form of visibility function can be obtained, which is used to assist the synthesis of novel viewpoint images, and can also be used for iterative optimization of depth maps.

This method first estimates an initial depth map using multi-view stereo^[40]. Then for each input viewpoint, it constructs a number of layers of depth planes parallel to the image plane to represent different depth values, and obtains a three-dimensional structure in space to express the depth distribution of this image. Each pixel (x, y) records a vote-value at the depth plane z as the possibility of existence of an object at the spatial position of depth z observed from (x, y) , while recording another value called vote-confidence to represent the credibility of vote-value. The initial vote-value and vote-confidence can be obtained simply using the initial depth map. Then the depth values of the adjacent viewpoint is used to calculate a consistent depth distribution $Consensus(x, y, z)$ considering the adjacent viewpoint information for each viewpoint. After obtaining consistent depth distribution, we can calculate the visibility information of each point.

$$SoftVis(x, y, z) = \max\left(0, 1 - \sum Consensus(x, y, z)\right)$$

Finally, the depth distribution and the visibility information of each image can be used for rendering the novel viewpoint image. Note that the visibility information can be used in the process of stereo matching to help calculating more accurate depth values. So the depth map can be optimized iteratively with visibility information, so that we can get depth map with higher quality.

This method fully considers the uncertainty of depth values, which includes the inconsistency of depth between adjacent pixels in the same image, and also includes the inconsistency between adjacent viewpoint images. The uncertainty between neighboring pixels within a single image is incorporated by filtering^[41,42], and the uncertainty across different views is incorporated using the soft reconstruction structure. Through this soft 3D reconstruction, a fixed depth value is not directly assigned to a certain pixel. Instead, the depth information of adjacent pixels and adjacent viewpoints is combined, and the inconsistency of depth information is well considered, so that good rendering results are obtained. In addition, the iterative optimization method of visibility and depth greatly improves the accuracy of depth value calculation at the boundary of objects in the image. Figure 10 shows the difference between the rendering result with a certain single depth value and the rendering result with soft 3D reconstruction. It can be seen that this method can deal with the problem caused by inaccurate depth values very well.

Figure 10 (a) Synthesized view using single depth; (b) Synthesized view using soft 3d reconstruction^[23].

The main disadvantage of this method is that since the depth value distribution of the image is expressed in a layered manner, the required storage space increases as the depth accuracy requirement increases. If the depth accuracy is not enough, when rendering a viewpoint far from the input image viewpoints, obvious error will occur. Therefore, a balance between storage space and depth precision is needed.

3.3.3 Rendering for casual photography

Hedman et al. present an algorithm which can reconstruct a 3D photo given a set of input images captured with a hand-held cell phone or DSLR camera, so that users can explore the scene freely^[43]. This allows the

data acquisition process to be independent of professional equipment and trained professionals, and even inexperienced average users can capture photos for 3D panoramic photo reconstructions with inexpensive cameras (Figure 11). Since the shooting is non-professional, there may be geometric distortion of the object in the scene or slight movement of the object between the acquired images, resulting in inconsistency of the three-dimensional information.

Figure 11 Capturing images using cell phone camera^[43].

Like other methods, this method first uses SFM to reconstruct the camera's parameters and the scene geometry represented using the sparse point cloud^[44]. Then complete depth maps need to be calculated using multi-view stereo. Because the photos is captured while moving the hand-held camera on a sphere of about half arm's length radius, traditional methods for depth estimation cannot obtain depth map with high quality due to the inconsistency. So they proposed a novel prior called near envelope that constrains the depth using a conservatively but tightly estimated lower bound, which highly improves the reconstruction. The obtained depth images need to be warped into the panoramic domain. For different pixels projected to the same position, the front point is selected by depth test to form the front surface of the panorama. Since the occluded content is important too, the back surface of the panorama is obtained by inverting the depth test. Finally, the front and back surfaces are merged into a uniform two-layer representation, resulting in a lightweight representation of the scene, which makes users explore freely. In addition, the method estimates the normal of the scene to enable some lighting effects.

This method can accommodate the inconsistency caused by the capturing process and improve the quality of panorama reconstruction. But the disadvantages are also obvious, that is, the time consumption is very large. It takes several hours to reconstruct a scene. Therefore, Hedman et al. proposed a new method called instant 3D photography^[45]. They jointly estimate the camera poses as well as spatially-varying adjustment maps that are applied to deform the depth maps and bring them into good alignment. They remove the need for label smoothness optimization, and replace it with independently optimizing every pixel label after filtering the data term in a depth-guided edge-aware manner, achieving two orders of magnitude speedup. Finally, they convert the 3D panorama into a multi-layered 3D mesh that can be rendered with standard graphics engines.

This kind of methods deal well with the inconsistency caused by casual capturing, making image-based rendering techniques not limited to professional capturing. But they are not suitable for rendering of large scale scenes.

3.3.4 Rendering combining global and local geometry

The rendering of indoor scenes is significantly different from outdoor scenes such as streetscapes. First of all, the spatial relationship of objects in indoor scenes is more complicated, and the complex occlusion relationship between objects needs to be considered carefully during the rendering process. In addition, indoor scenes are mostly man-made objects, and due to the material properties of these objects, different textures can be observed when viewed from different locations. This results in a large inconsistency in obtaining scene visual information from multiple images of different perspectives. Even if we can reconstruct the global texture, we will lose the difference from different angles, which will greatly affect the realism. In order to deal with the difficulty of indoor scene rendering, Hedman et al. proposed a new image-based rendering method^[46], combining the global and local information of the scene, to obtain a more realistic rendering effect (Figure 12).

This method takes as input the RGBD image taken by the consumer-level depth sensor. It first aligns the

Figure 12 Rendered images using method combining global and local geometry^[46].

input image using SFM and reconstructs the global geometric information of the scene. Local geometric information is then created for each input viewpoint such that the local geometric information is aligned with the edges in the image. Next, the multi-view stereo method guided by the global geometric information of the scene is used to calculate the depth map for each viewpoint, and convert these depth maps into a simplified mesh representation suitable for rendering. Finally, the tile structure is used to store these simplified meshes to improve rendering efficiency. When rendering, it firstly finds the local geometric information related to the new viewpoint to be rendered according to the tile data structure, and projects it into the new viewpoint image for fusion. So the rendering result of the new viewpoint is obtained.

For indoor scenes, inconsistency in different viewpoints needs to be preserved. This approach sacrifices part of the global property by using the local geometric information of each viewpoint to respect different viewpoints. This allows users to observe view-dependent appearance such as highlights during the roaming process and to observe more accurate scene geometry. However, the calculation is relatively expensive, so it is difficult to handle large scenes with too many images. In addition, it requires a depth sensor, which is only suitable for indoor scene applications, but is not applicable to outdoor scenes.

4 Conclusions

According to whether the three-dimensional information of the scene is utilized, the image-based rendering techniques can be divided into two categories. Among them, the methods of image-based rendering without using three-dimensional information is mostly early research results. They generally obtain the image of the novel viewpoint using image stitching, and do not need to reconstruct the geometric structure. But they often require that the cameras have a specific layout, and they can only provide free-view roaming within a certain range. The cylindrical panorama and the spherical panorama can make users explore freely at a fixed position, while the concentric mosaic and the spherical light field capture system can make the observation point freely move within a certain range.

In recent years, people have paid more attention to the use of three-dimensional information to help image-based rendering. We first introduce the methods implicitly utilizing the three-dimensional information of the scene, and then we introduce the methods explicitly using the three-dimensional information of the scene. The so-called implicit use of three-dimensional information means the use of image correspondence when performing viewpoint interpolation. Although this kind of methods do not explicitly reconstruct the 3D model of the scene, since the correspondence between images actually contains the geometric information, we divide it into categories that utilize the three-dimensional information of the scene. Among the methods using image correspondence, the method of Chen and Williams is only applicable to the case where the image baseline is narrow^[24], while the method of Nie et al.

is proposed to deal with the wide baseline case^[26].

The methods explicitly utilizing the three-dimensional information of the scene use explicit representation of the scene geometry obtained by reconstructing the depth map, the three-dimensional point cloud or the triangle mesh to help obtaining high-quality rendering results. Such methods are further divided into methods focusing on the lack of geometric information and methods focusing on the consistency of geometric information. The methods focusing on the lack of geometry complement the missing area using the geometric information of adjacent area, which can compensate for the inaccurate recovery of geometry induced by the capturing process etc. Paying attention to the consistency of geometric information is a research direction in recent years, which includes not only reducing inconsistency caused by capturing, but also retaining the inconsistency that should be retained between different images. Consistent correction of inconsistent depth estimation is done by methods based on image correspondence and depth map to avoid ghosting in rendering result. Rendering methods for casual photography tolerate poor depth reconstruction by tolerating geometric inconsistencies between images. The rendering method based on soft 3D reconstruction and the rendering method combining global and local geometric information considers the consistency of global geometric information while considering the inconsistency of geometric information under different viewpoints, so that they can improve the realism of rendering. A comparison of several representative methods is shown in Table 1.

Table 1 Comparison of several representative methods

	Geometric information	Constraint of the input viewpoints	Characteristic
Cylindrical panorama	None	Strict	Low computational complexity, only for views from a single point
Concentric mosaic	None	Strict	A specific region can be explored freely
Spherical light field	None	Strict	A specific region can be explored freely
Chen et al. 1993	Implicitly	Free	Weak constraint on input views
Nie et al. 2017	Implicitly	Free	Dealing with wide-baseline images
Chaurasia et al. 2013	Explicitly	Free	Tolerance of inaccurate 3D information
Lipski et al. 2014	Explicitly	Free	Tolerance of inconsistent 3D information
Penner et al. 2017	Explicitly	Free	Modeling the uncertainty of 3D information
Hedman et al. 2017	Explicitly	Free	Tolerance of rough input images with inconsistent 3D information
Hedman et al. 2018	Explicitly	Free	Tolerance of rough input images; high computational efficiency
Hedman et al. 2016	Explicitly	Free	Modeling the uncertainty; suitable for indoor scenes

In fact, with the development of extracting three-dimensional geometric information from images, image-based rendering using three-dimensional information has become more popular, and has a wider application. Most image-based rendering methods are multi-step systematic approaches. People should not only pay attention to the method of interpolation of novel viewpoint images, but also pay attention to the acquisition of geometric information such as depth map and image correspondence, the consistent representation between three-dimensional information and images, and the post-processing process after interpolation. The development of many methods in other areas have also played an important role in promoting image-based rendering techniques. For example, there may be many holes in the novel viewpoint image obtained by many image-based rendering methods. In this case, we need to use the image completion methods for post processing^[47]. When rendering a novel view image, the pixels' color of the novel image are often from different input images, which means we should blend them for the final rendering result^[48].

In this paper, we only discuss methods using multi-view geometry as connection of the 2D information of images and 3D information of the scene. However, in recent years, with the rapid development of deep learning, deep learning based methods also achieve good performance^[49,50]. Unlike methods we mentioned

above, it provides end-to-end rendering process, that is, directly rendering the image of novel view from the discretely sampled input images. For example, DeepStereo uses a large number of real scene photos as training set to train a depth prediction network and a color prediction network, so that people can obtain rendering result directly from input pixels^[50]. This method can deal with the wide-baseline case, and due to the ability of deep network to fit very complex nonlinear functions, some problems which traditional methods are difficult to handle can be better processed.

In recent years, researchers have paid more attention to how to deal with the problem of inaccurate restoration of geometric information caused by capturing or the property of the scene. For example, the geometric information of small or thin objects is difficult to recover, and the images have poor consistency when capturing dynamic scenes. These open problems are interesting future works. In addition, with the development of deep learning, the use of deep learning techniques to deal with some of the problems of traditional methods will also be an interesting future work.

References

- 1 Snavely N, Seitz S M, Szeliski R. Photo Tourism: Exploring image collections in 3D. *ACM Transactions on Graphics*, 2006, 25: 835–846
DOI:10.1145/1141911.1141964
- 2 Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010, 32(8): 1362–1376
DOI:10.1109/TPAMI.2009.161
- 3 Google Street View. <https://www.google.com/streetview/>
- 4 Shum H Y, Kang S B. A review of image-based rendering techniques. In: Ngan K N, Sikora T, Sun M T, eds. *Proceedings of IEEE/SPIE Visual Communications and Image Processing (VCIP)*, Perth. Australia, 2000, 2–13
- 5 Adelson E H, Bergen J R. The plenoptic function and the elements of early vision. In: Landy M S, Movshon J A, eds. *Computational Models of Visual Processing*. Cambridge: MIT Press, 1991
- 6 Wong T T, Heng P A, Or S H, Ng W Y. Image-based rendering with controllable illumination. In: Dorsey J, Slusallek P, eds. *Proceedings of the 8-th Eurographics Workshop on Rendering*. Berlin: Springer-Verlag, 1997, 13–22
- 7 McMillan L, Bishop G. Plenoptic modeling: An image-based rendering system. In: Mair SG, Cook R, eds. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York: ACM, 1995, 39–46
- 8 Levoy M, Hanrahan P. Light field rendering. In: Fujii J, eds. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. New York: ACM, 1996, 31–42
DOI:10.1145/237170.237199
- 9 Shum H Y, He L W. Rendering with concentric mosaics. In: Waggenspack W, eds. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. New York: ACM Press/Addison-Wesley Publishing Co., 1999, 299–306
DOI:10.1145/311535.311573
- 10 Szeliski R, Shum H Y. Creating full view panoramic image mosaics and texture-mapped models. In: Owen G S, Whitted T, Mones-Hattal B, eds. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. New York: ACM Press/Addison-Wesley Publishing Co., New York, 1997, 251–258
DOI:10.1145/258734.258861
- 11 Chen S E. Quick Time VR—an image-based approach to virtual environment navigation. In: Mair SG, Cook R, eds. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York: ACM, 1995, 29–38
DOI:10.1145/218380.218395
- 12 Roundshot. <http://www.roundshot.com>
- 13 Shum H Y, Szeliski R. Construction and refinement of panoramic mosaics with global and local alignment. In: *Sixth International Conference on Computer Vision (ICCV'98)*. Bombay, 1998, 953–958
DOI:10.1109/ICCV.1998.710831
- 14 Gortler S J, Grzeszczuk R, Szeliski R, Cohen M F. The lumigraph. In: Fujii J, eds. *Proceedings of the 23rd annual*

- conference on Computer graphics and interactive techniques. New York: ACM, 1996, 43–54
DOI:10.1145/237170.237200
- 15 Debevec P, Downing G, Bolas M, Peng H-Y, Urbach J. Spherical light field environment capture for virtual reality using a motorized pan/tilt head and offset camera. In: ACM Siggraph 2015 Posters. Los Angeles, California, 2015, 1–1
DOI:10.1145/2787626.2787648
 - 16 Buehler C, Bosse M, McMillan L, Gortler S, Cohen M. Unstructured lumigraph rendering. In: Pocock L, eds. Proceedings of the 28th annual conference on Computer graphics and interactive techniques. New York: ACM, 2001, 425–432
DOI:10.1145/383259.383309
 - 17 Rademacher P. View-dependent geometry. In: Waggenspack W, eds. Proceedings of the 26th annual conference on Computer graphics and interactive techniques. New York: ACM Press/Addison-Wesley Publishing Co., 1999, 439–446
DOI:10.1145/311535.311612
 - 18 Vedula S, Baker S, Kanade T. Spatio-temporal view interpolation. In: Gibson S, Debevec P, eds. Proceedings of the 13th Eurographics Workshop on Rendering. Aire-la-Ville: Eurographics Association, 2002
 - 19 Chaurasia G, Duchene S, Sorkine-Hornung O, Drettakis G. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics*, 2013, 32(3): 1–12
DOI:10.1145/2487228.2487238
 - 20 Lipski C, Klose F, Magnor M. Correspondence and depth-image based rendering a hybrid approach for free-viewpoint video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, 24(6): 942–951
DOI:10.1109/TCSVT.2014.2302379
 - 21 Shade J, Gortler S, He L W, Szeliski R. Layered depth images. In: Cunningham S, Bransford W, Cohen M F, eds. Proceedings of the 25th annual conference on Computer graphics and interactive techniques. New York: ACM, 1998, 231–242.
DOI:10.1145/280814.280882
 - 22 Chang C, Bishop G, Lastra A. LDI tree: A hierarchical representation for image-based rendering. In: Waggenspack W, eds. Proceedings of the 26th annual conference on Computer graphics and interactive techniques. New York: ACM Press/Addison-Wesley Publishing Co., 1999: 291–298
DOI:10.1145/311535.311571
 - 23 Penner E, Zhang L. Soft 3D reconstruction for view synthesis. *ACM Transactions on Graphics*, 2017, 36(6): 1–11
DOI:10.1145/3130800.3130855
 - 24 Chen S, Williams L. View interpolation for image synthesis. In: Whitton MC, eds. Proceedings of the 20th annual conference on Computer graphics and interactive techniques. New York: ACM, 1993, 279–288
DOI:10.1145/166117.166153
 - 25 Seitz S M, Dyer C M. View morphing. In: Fujii J, eds. Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. New York: ACM, 1996, 21–30
DOI:10.1145/237170.237196
 - 26 Nie Y W, Zhang Z S, Sun H Q, Su T, Li G. Homography propagation and optimization for wide-baseline street image interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 2017, 23(10): 2328–2341
DOI:10.1109/TVCG.2016.2618878
 - 27 Kopf J, Langguth F, Scharstein D, Szeliski R, Goesele M. Image-based rendering in the gradient domain. *ACM Transactions on Graphics*, 2013, 32(6): 1–9
DOI:10.1145/2508363.2508369
 - 28 Sinha S N, Kopf J, Goesele M, Scharstein D. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics*, 2012, 31(4): 1–10
 - 29 Thonat T, Djelouah A, Durand F, Drettakis G. Thin structures in image based rendering. *Computer Graphics Forum*, 2018, 37: 107–118
 - 30 Yang H S, Lin W Y, Lu J B. DAISY Filter Flow: A generalized discrete approach to dense correspondences. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, 2014, 3406–3413
 - 31 Bao L C, Yang Q X, Jin H L. Fast edge-preserving PatchMatch for large displacement optical flow. *IEEE Transactions on Image Processing*, 2014, 23: 4996–5006
DOI:10.1109/CVPR.2014.452
 - 32 Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(11): 2274–2282

- DOI:10.1109/TPAMI.2012.120
- 33 Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM*, 1981, 24(6): 381–395
DOI:10.1145/358669.358692
- 34 Hawe S, Kleinstaub M, Diepold K. Dense disparity maps from sparse disparity measurements. In: 2011 International Conference on Computer Vision. Barcelona, 2011, 2126–2133
DOI:10.1109/ICCV.2011.6126488
- 35 Sinha S N, Steedly D, Szeliski R. Piecewise planar stereo for image-based rendering. In: 2009 IEEE 12th International Conference on Computer Vision. Kyoto, 2009, 1881–1888
- 36 Gallup D, Frahm J, Pollefeys M. Piecewise planar and non-planar stereo for urban scene reconstruction. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, 2010, 1418–1425
- 37 Furukawa Y, Curless B, Seitz S M, Szeliski R. Manhattan-world stereo. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, 2009, 1422–1429
DOI:10.1109/CVPR.2009.5206867
- 38 Lipski C, Linz C, Neumann T, Wacker M, Magnor M. High resolution image correspondences for video post-production. In: 2010 Conference on Visual Media Production. London, 2010, 33–39
DOI:10.1109/CVMP.2010.12
- 39 Liu C, Yuen J, Torralba A. SIFT flow: Dense correspondence across different scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(5): 978–994
DOI:10.1109/TPAMI.2010.147
- 40 Hosni A, Bleyer M, Rhemann C, Gelautz M, Rother C. Real-time local stereo matching using guided image filtering. In: 2011 IEEE International Conference on Multimedia and Expo. Barcelona, 2011, 1–6
DOI:10.1109/ICME.2011.6012131
- 41 Ma Z Y, He K M, Wei Y C. Constant time weighted median filtering for stereo matching and beyond. In: 2013 IEEE International Conference on Computer Vision. Sydney, NSW, 2013, 49–56
- 42 He K M, Sun J, Tang X O. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013, 35: 1397–1409
DOI:10.1109/TPAMI.2012.213
- 43 Hedman P, Alsian S, Szeliski R, Kopf J. Casual 3D photography. *ACM Transactions on Graphics*, 2017, 36: 234: 1–15
- 44 Schönberger J L, Zheng E L, Frahm J M. Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision. Amsterdam, the Netherlands, 2016
- 45 Hedman P, Kopf J. Instant 3D photography. *ACM Transactions on Graphics*, 2017, 36(6): 1–15
DOI:10.1145/3130800.3130828
- 46 Hedman P, Ritschel T, Drettakis G, Brostow G. Scalable inside-out image-based rendering. *ACM Transactions on Graphics*, 2016, 35(6): 1–11.
DOI:10.1145/2980179.2982420
- 47 Pérez P, Gangnet M, Blake A. Poisson image editing. *ACM Transactions on Graphics*, 2003, 22: 313–318
DOI:10.1145/1201775.882269
- 48 Burt P J, Adelson E H. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics*, 1983, 2: 217–236
DOI:10.1145/245.247
- 49 Kalantari N K, Wang T C, Ramamoorthi R. Learning based view synthesis for light field cameras. *ACM Transactions on Graphics*, 2016, 35(6): 1–10
DOI:10.1145/2980179.2980251
- 50 Flynn J, Neulander I, Philbin J, Snavely N. DeepStereo: Learning to predict new views from the world’s imagery. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, 2016, 35(6): 1–10