RGB-D IBR: Rendering Indoor Scenes Using Sparse RGB-D Images with Local Alignments

Yeongyu Jeong
POSTECH

Haejoon Kim
POSTECH

Hyewon Seo
University of Strasbourg

Frederic Cordier
University of Haute-Alsace

Seungyong Lee
POSTECH

(a) part of the input image set (b) rendered image based on (a) (c) corresponding real image

Figure 1: *Our method uses sparse RGB-D images as input and generates images seen from novel views. Note that (c) is not involved in the blending for synthesizing (b).*

Abstract

This paper presents an image-based rendering (IBR) system based on RGB-D images. The input of our system consists of RGB-D images captured at sparse locations in the scene and can be expanded by adding new RGB-D images. The sparsity of RGB-D images increases the usability of our system as the user need not capture a RGB-D image stream in a single shot, which may require careful planning for a hand-held camera. Our system begins with a single RGB-D image and images are incrementally added one by one. For each newly added image, a batch process is performed to align it with previously added images. The process does not include a global alignment step, such as bundle adjustment, and can be completed quickly by computing only local alignments of RGB-D images. Aligned images are represented as a graph, where each node is an input image and an edge contains relative pose information between nodes. A novel view image is rendered by picking the nearest input as the reference image and then blending the neighboring images based on depth information in real time. Experimental results with indoor scenes using Microsoft Kinect demonstrate that our system can synthesize high quality novel view images from a sparse set of RGB-D images.

Keywords: image-based rendering, RGB-D images, local alignment, 3D navigation

Concepts: •**Computing methodologies** → **Image-based rendering;**

I3D '16, February 27-28, 2016, Redmond, WA, USA
ISBN: 978-1-4503-4043-4/16/02
DOI: http://dx.doi.org/10.1145/2856400.2876006

1 Introduction

Image-based rendering (IBR) enables 3D navigation of a scene by generating novel view images from a set of images captured in the scene. Novel view synthesis can be achieved by interpolation or warping of input images. Previous approaches use RGB images only, and geometric information used for interpolation or warping should be reconstructed by analyzing the input images. Geometric information from a few RGB images could be inaccurate, and novel views may not be generated far from the views of input images. For plausible 3D navigation, a dense set of images should be captured from the scene.

Nowadays depth cameras, such as Microsoft Kinect, have become widely available and the geometric information as well as RGB values of an image can readily be obtained. This additional depth information can be used to estimate accurate 3D geometry of the scene, which can be utilized to synthesize more accurate novel views even with a smaller number of input images. In this paper, we investigate on image-based rendering for 3D navigation using RGB-D images that have been captured at sparse locations in a scene with a hand-held camera.

Our system takes a sparse set of RGB-D images as input and allows the user to capture the input RGB-D images in a convenient way. The user can take only desirable images whenever needed, and does not have to hold a camera all the time and to struggle with image blurs caused by camera motion. In addition, no spatial coherence is assumed between input images, and there is no constraint on the order and the trajectory of image capture. However, with sparse capture locations, input images would have small overlaps and large visibility changes, and the alignments between them could become less reliable. To overcome this problem, we use geometric information that can be induced from depth values of input RGB-D images.

In addition to the design and implementation of an image-based rendering system based on a sparse set of RGB-D images with local alignments, our contributions can be summarized as follows;

- We present an efficient and effective method for localizing an

Figure 2: *Result images of Office (top) and Lounge (bottom) scenes. The first and third columns show rendered images, while the second and fourth columns show real images taken from similar views to the synthesized ones.*

input image with respect to the whole input images, handling the ambiguity due to repeated structures of indoor scenes.

- We improve Affine SIFT matching [Yu and Morel 2011] for fast and accurate alignments of RGB-D images, utilizing the depth information to reduce the number of viewpoint pairs.
- We present a real-time view synthesis algorithm for a local alignment graph of RGB-D images, enhancing the visibility test and blending weight computation for input RGB pixels.

2 System Overview

Our system takes a sparse set of RGB-D images as input and generates an image rendered from an arbitrary viewpoint. Input images are added incrementally and there is no ordering constraint between them. We assume that input image pairs have wide baselines with small overlapping areas compared to video frames. Input images are maintained as a graph, where each node is an image and an edge contains relative pose information between nearby images. Information about neighbor images consists of a list of nearby images and relative camera poses for them. When an image is newly added, the graph can be updated within interactive time, and novel view images can be rendered in real time using the graph.

Our system consists of three steps: localization, alignment, and rendering. The localization step takes a new input image and locates it relative to the existing images in the system. Neighbor images containing partial overlaps with the new image are identified in this step. The alignment step calculates relative pose information between the new image and its neighbors. The rendering step generates novel view output images and enables the user to explore the scene in real time. For the localization step, we extract visual features from each image and use them for finding neighbor images. In the alignment step, relative camera poses are computed by our improved version of Affine SIFT (ASIFT) [Yu and Morel 2011]. In the rendering step, an intermediate depth image for a novel view is constructed using the depth information of the neighbor images and used for blending RGB pixel values.

The noticeable advantages of our system compared to other methods are as follows. First, our system runs robustly with a sparse set of RGB-D images, where most of previous works rely on spatial coherency of video streams. Next, we improve the speed of feature matching in the alignment step by utilizing the geometric information of depth images. Finally, we avoid time-consuming global alignment of input images, and need not suffer from maintaining a single geometric model for the whole scene.

3 Results

We implemented our proposed system using C++, and CUDA was used for normal map filtering of RGB-D images. The rendering step was implemented using fragment shader. We used Microsoft Kinect 1 for capturing RGB-D images, and our system was experimented with indoor scenes running on Intel Core i7 920 @ 2.67 GHz, 6GB RAM, NVIDIA GeForce GTX 680.

The rendering results of our system are shown in Fig. 2. The most time-consuming part is the alignment step, and it requires more computation for a denser set of input images which would have more neighbors for each image. Office and Lounge scenes in Fig. 2 use three neighbor images for most input images, while more neighbors are needed for a scene with many occluders. Every process inside the rendering step is performed for each frame in real time. Our system provides a synthesized novel-view image with more than 60 fps which should be satisfactory for navigating a virtual scene. Fig. 2 shows that synthesized images are quite similar to the corresponding real ones, where the real images are not used for synthesizing the novel-view images.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant (NRF-2014R1A2A1A11052779) and Institute for Information & communications Technology Promotion (IITP) grant (R0126-15-1078) both funded by the Korea government (MSIP).

References

YU, G., AND MOREL, J.-M. 2011. ASIFT: An algorithm for fully affine invariant comparison. *Image Processing On Line 1*.