Room Reconstruction with Microsoft kinect

Mattia Bonomi, Federico Zanetti

Trento, 15/03/2015

The Kinect

Kinect is based on software technology developed by Microsoft Game Studios owned by Microsoft and on range camera technology by israeli developer PrimeSense.
The depth sensor outputs on a specific stream the depth map constructed by analyzing a speckle pattern of infrared laser light originated by a specific IR transmitter. The Kinect uses an infrared projector and sensor; it does not use its RGB camera for depth computation.
The depth computation is all done by the PrimeSense hardware built into Kinect. The depth map is constructed by analyzing the predetermined pattern of infrared dots. Such technique of analyzing a known pattern is called structured light and its general principle is to infer the depth from the deformation of the projected pattern.

The Kinect combines structured light with two classic computer vision techniques: depth from focus, and depth from stereo. The Kinect dramatically improves the accuracy of traditional depth from focus. The Kinect uses a special ("astigmatic") lens with different focal length in x and y directions.
The astigmatic lens causes a projected circle to become an ellipse whose orientation depends on depth.
Depth information from the stereo vision uses parallax. The sensor is placed at an offset relative to the IR transmitter, and the difference between the observed and expected IR dot positions is used to calculate the depth at each pixel of the RGB camera. If the scene is looked at from another angle, objects that are closer get shifted to the side more than those that are farther away. The Kinect analyzes the shift of the speckle pattern by projecting from one location and observing from another.

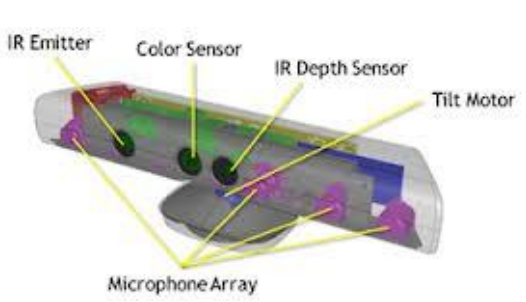


Figure 1: basic structure of the Kinect.

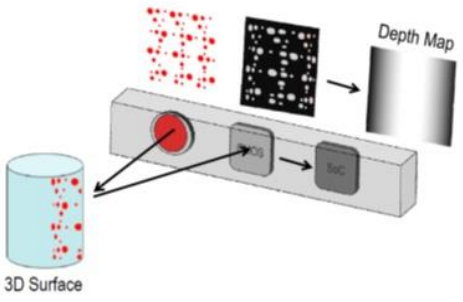


Figure 2: working scheme of IR transmitter and IR sensor.

Parameters

The kinect device includes three foundamenal elements: a color VGA video camera, a depth sensor and a multi-array microphone. A look at the Kinect for windows SDK technical specifications reveal that both the video and depth sensor cameras have a 640 x 480-pixel resolution and run at 30 frames per second.

The working range lies between 80 cm and 400 cm with a depth estimation error of 1 DN at the most at all depths. Since DN to depth convesion is non linear this is equivalent to 1 mm at closest distance and of 5 mm at farthest distance.
The values for the field of view given by the API are the following: for the color camera the horizontal FOV is 62,0°and the vertical FOV is 48,6°. For the depth camera instead the horizontal FOV: 58,5°and the horizontal FOV: 45,6°.

Future works

Capturing a full color 3D model of an indoor space and creating 3D meshes out of real scene opens the doors to new applications. Once the virtual model reconstruction procedure is fully defined, it may be significant to carry out modifications to the model such as object disposal or object insertion and affine manipulations. Even the automatic identification of a set of objects in 3D meshes has several applications.

As stated in [xy], with object segmentation procedures through direct user interaction users may also wish to scan a specific smaller physical object rather than the entire scene. In order to do this, the user may reconstruct the entire scene, and then accurately segment the desired object. The system shall monitor the 3D reconstruction and observe changes between different acquisitions. If an object is physically removed from view or moved within the scene by the user, large detectable changes in the 3D virtual model are observed. When such changes are perceived, the repositioned object shall be cleanly segmented from the background model. This approach allows a user to perform segmentation without any explicit GUI input, simply by moving the object directly.

Being able to intelligently suggest object bounding boxes could shall be useful in order to improve manipulations to the meshes and the 3D scene understanding. It may even support a robot with a mounted sensor to navigate its environment and autonomously acquire a database of objects from its surroundings without being explicitly presented every object. Bounding boxes estimation from room reconstruction may facilitate proper 3D design for furniture and indoor design.

Ref [xy]

http://research.microsoft.com/pubs/155416/kinectfusion-uist-comp.pdf

KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera*
Shahram Izadi1, David Kim1,3, Otmar Hilliges1, David Molyneaux1,4, Richard Newcombe2, Pushmeet Kohli1, Jamie Shotton1, Steve Hodges1, Dustin Freeman1,5, Andrew Davison2, Andrew Fitzgibbon1

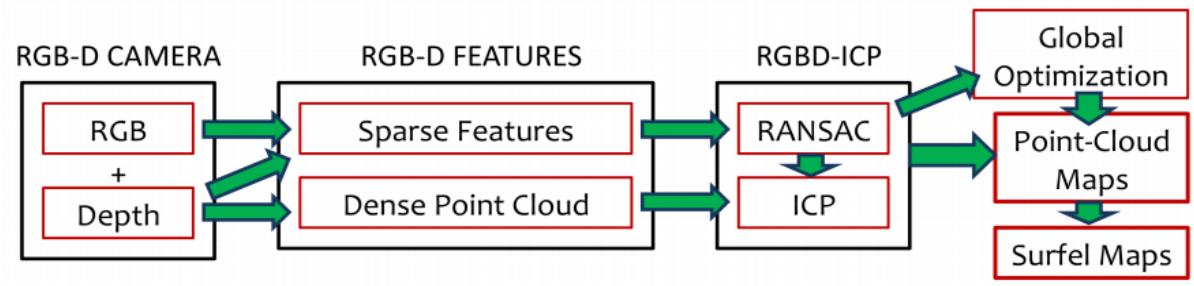
Main weaknesses

We recommend a practical procedure to secure a room scanning of quality. The core system described so far makes assumptions that the scene will remain reasonably static, clearly to avoid any confusion between adjacent point clouds. The system is all the same noise-sensitive and the acquisition turns out to be demanding in terms of number of frames and time in order to drastically reduce disturbances. To perform 180° or even 360° acquisition, we spun the kinect by very little angle at a time (roughly 5°).

Given the kinect parameters and the inherent error of the depth measurement small details are hardly detectable and the effect explodes at the farther depth range limit. For a neat noise-free acquisition small objects should be removed by hand in advance or by the implemented filters.

Another nasty effect may occur as a 180° acquisition is performed once a third room wall is detected. As the kinect moves along surfaces whose depth varies gradually with respect to kinect's point of view and the field of view is too little to include a reference element in the scene, the reconstruction may lose track of the correct displacement of the walls (scene's most relevant structural element). To avoid such inconvenient effect, it is compelling to set a wider field of view possibly including a unifying element which for practical purposes may be found in the floor. This imposes to work on a deeper map which may more easily bring in depth errors, but the overall result is more appraisable.

Figura 1: wrong overlap occurring on close proximity to the scene


RGB-D CAMERA
RGB
+
Depth
RGB-D FEATURES
Sparse Features
Dense Point Cloud
RGBD-ICP
RANSAC
ICP
Global Optimization
Point-Cloud Maps
Surfel Maps