Kinect Unleashed: Getting Control over High Resolution Depth Maps

Manuel Martinez and Rainer Stiefelhagen
Institute of Anthropomatics, Karlsruhe Institute of Technology
Karlsruhe, Germany
manuel.martinez@kit.edu and rainer.stiefelhagen@kit.edu

Abstract

The wide availability and economic price of Kinect popularized the concept of RGB-D cameras, which are increasingly used in a wide range of applications ranging from Human Computer Interfaces to Structural Analysis, Health Care and even Art. With over 20 million units sold, it is by far the most popular depth camera, although, being closed source, its internals are not entirely known. In this paper we focus on the PS1080 chip that executes the depth extraction algorithm inside Kinect and related devices. We propose an algorithm that can estimate the output of Kinect from the raw infrared camera data. Using this algorithm, the internal reference calibration image from Kinect can be obtained. This image allows us to model Kinect as a standard stereo camera, where alternative stereo algorithms can be used increasing the versatility of Kinect.

1 Introduction

Before Kinect, the most used devices to capture a robust depth field were Laser Scanners and Time-Of-Flight cameras, but the cost of those devices is overwhelmingly high and they are seldomly used besides specific industrial tasks. Most domestic users were limited to stereo cameras which are not as reliable for the task.

Even though Kinect was developed to be used with the XBOX360 console, due to the high expectations raised it was hacked within hours of its release and made available to the public in general. A wide variety of applications have followed up in all kind of fields.

Although the high popularity of Kinect as the *de facto* RGB-D device, it is still a hacked, closed source device designed for a very specific task: consumer-level Human Computer Interaction. As such it has two strong limitations:

- Its specifications are not stated by the manufacturer but estimated by several research works.
- It uses a fixed depth algorithm that can not be altered to fit the specifics of the task.

The computer vision subsection of Kinect was developed by PrimeSense. It comprises an infrared laser projector at 830nm, an infrared camera with a matching bandpass filter, and a PS1080 chip that processes the output from the IR camera to calculate the depth map. The PS1080 supports optionally a RGB camera, and has the capability of registering the depth field to the RGB camera to obtain a RGB-D output. This capability is disabled on Kinect, but it is available in other devices based on the same hardware such as Wavi Xtation and the PrimeSense Reference Device.

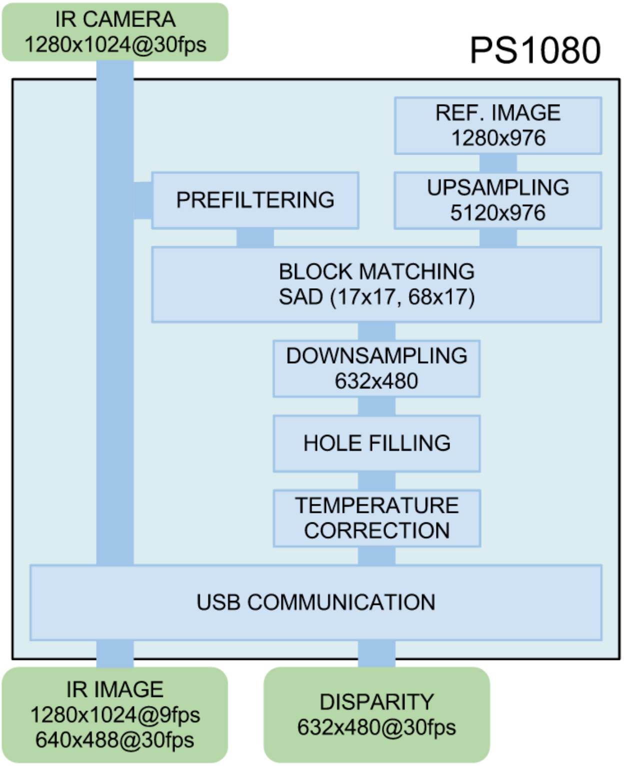

Figure 1. Our suggested architecture to emulate PS1080. The PS1080 acts as a stereo camera where one of the sensors has been replaced by a fixed pattern stored in the internal memory. Using a fixed pattern can create large shift in albedo which is compensated by a prefiltering step, and a postfiltering step is added to compensate for temperature changes.

Several works have analyzed the Kinect as a whole system. Andersen *et al.* [1] published one of the most comprehensive technical reports about Kinect from the CV field, citing a large amount of characteristics of Kinect. Other works have focused on modelling the Kinect as a Range Finding device [7, 10] analyzing in detail the generated depth field.

Few works however have focused on modeling the actual depth process that is performed inside Kinect. A detailed forensic analysis, of Kinect depth field has been performed team behind ROS [9]. This work highlight the similarity between Kinect and alternate projected texture approaches [8].

We will further analyze the depth process inside Kinect by modeling the algorithm inside the PS1080 chip. The novelty of our approach is that we use both the input and the output of the PS1080 chip, to model it as in a black box testing procedure.

We found that most calibration information is stored as a reference image, and we show a procedure to approximate it. This procedure takes less than 80 seconds and must be run once for each Kinect. Using the estimated reference image, our model can reproduce Kinect disparity map with an average error of 0.02 pixels.

Reproducing Kinect disparity map is useful in order to gain knowledge about its capabilities. But an alternate use of the reference image is to simulate a stereo camera system. Therefore all algorithms designed for stereo cameras can be used on Kinect, increasing its versatility.

2 PS1080 Description

The depth processing system of PS1080 is simple but effective, and it is based on triangulation like a standard stereo camera.

Stereo cameras extract depth by capturing two diferent views from the scene, finding the correspondences between them, and extract depth by triangulation. However, occlusions, repeated patterns and lack of texture difficults the correspondence problem in the general case.

Kinect simplifies the correspondence problem by projecting a texture over the scene. This texture is carefully designed to present unique patterns along the epipolar lines to enforce that the correspondence can be found simply by correlation. This approach is known as *projected texture* [8].

However, for consumer and domestic use, projecting a texture was not considered a viable option: it is invasive as it projects a strong texture in the scene, and it is expensive as it requires two cameras and one projector. The first problem was solved in Kinect by switching to infrared illumination. The second was solved by using a fixed pattern light source in detriment of spatial resolution and replacing one of the cameras by the structured light pattern itself.

2.1 Transmission Modes

The PS1080 in Kinect has three transmission channels:

- Control channel to send and receive configuration and status messages.
- Image channel, used to send either data from the RGB camera or the IR camera, but not both simultaneously. For RGB camera there are several multiple formats available. For IR, non-gamma corrected 11-bit precision data is sent at two possible resolutions: 640x488@30fps(VGA) and 1280x1024@9fps(SXGA).
- Depth channel, used to send disparity values at VGA@30fps. Format is fixed point with 8 bits for the integer part and 3 bits for the fractional part. It is only available when the IR camera is configured at 640x488@30fps.

The main limitation of this approach is the impossibility of sending IR and RGB images simultaneously, therefore it is not possible to use an alternate stereo demo together with RGB imaging. However, using an external RGB camera with the depth map of Kinect is not difficult and procedures to calibrate the system are known [4].

2.2 Projected Pattern

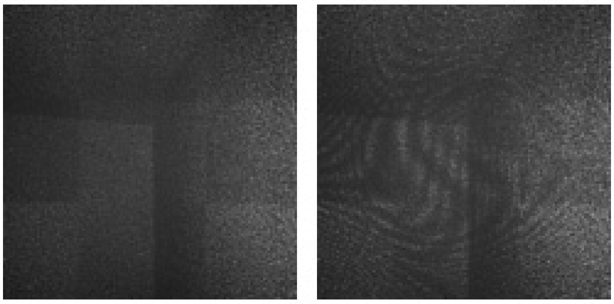

Figure 2. The projected IR pattern from kinect as seen from the IR camera. Left: sensor is configured at SXGA. Right: sensor is configured at VGA, and a moire pattern is present.

A 830nm, temperature stabilized laser goes through a first diffraction grating where it is divided in a 3x3 beams. A second diffraction grating splits each beam to form a pattern of 211x165 uncorrelated quasi-periodic dots. This pattern is the same for all Kinects.

Using an uncorrelated pattern helps in the correspondence problem and increases the robustness against interference. In a similar approach from 2001 device with a periodic pattern [2] each dot needed to be very far apart from the neighboring one in order not to be confused with it, thus limiting the available resolution.

The density of the dot pattern is enough to create a moiré pattern when recorded at VGA resolution (Fig. 2).

2.3 Kinect Depth Pipeline

We consider Kinect as a stereo camera whose 2nd camera has been substituted by an IR projector.

Furthermore, we known that Kinect is designed as a low cost depth camera whose processing is done in hardware. This, together and the results from several Kinect works, guides us into the formulating the following constraints:

- IR view from the camera is not rectified, as it is expensive to do it in hardware. A carefully crafted low-distortion lens is used instead.
- As most economic hardware based approaches, a Block Matching Sum of Absolute Diferences algorithm solves the correspondence problem.
- To reduce memory bandwidth, the epipolar lines are approximated by the horizontal scanlines.

Therefore, we suggest the approach shown in Fig. 1 as a possible architecture for the PS1080.

From the published Kinect driver from PrimeSense we know that a reference image exists and its size is

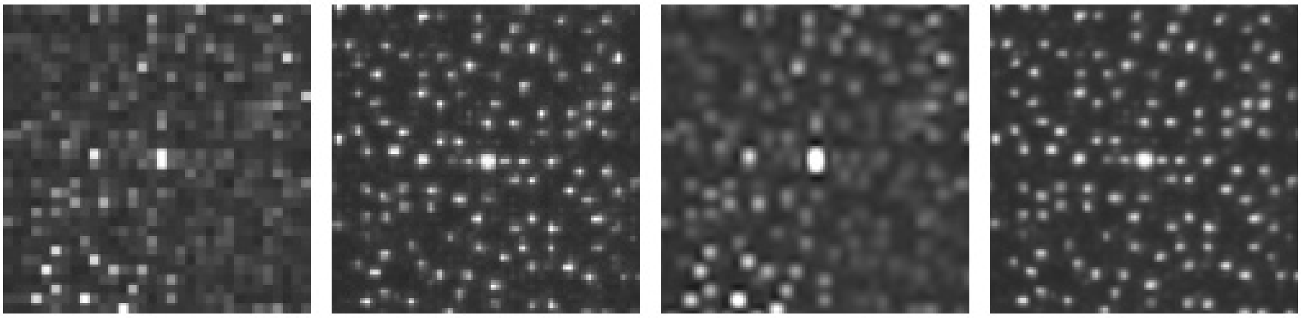

Figure 3. Detail of the reference image recovered from different sources. From left to right: VGA, SXGA, VGA+superresolution, SXGA+superresolution.

1280x1024. Moreover, the speed of the IR Image sensor is coherent with a 1280x1024@30fps. This information suggests that the depth extracting algorithm is performed internally at 1280x1024. However, Kinect provides depth extraction at 632x480 with 3 bits of subpixel precision. Usually subpixel precision is obtained using quadratic interpolation, but this is expensive in hardware. Block matching can be efficiently implemented instead, so we suggest that the internal reference image is upsampled at 5120x1024 internally and the SAD is done on an assimetrical window, where 17x17 pixels from the IR image are matched against 68x17 pixels of the reference.

3 Getting Reference Image

To approximate the reference image, we reverse the stereo camera model:

$$D(x,y) = \arg\min_{k} |I(x+k,y) - R(x,y)| \tag{1}$$

turns into into:

$$R'(x,y) = i(x + D(x,y), y) \ . \tag{2}$$

Therefore, if we obtain simultaneously the IR image I and the disparity field D, we can obtain an approximation of the reference image R' by displacing the pixels from I the amount noted by D.

In order for this to work, we assume that, for the optimal k, $I(x+k,y) \approx R(x,y)$. A more detailed model would assume:

$$I(x+k,y) \approx A(x+k,y)\ R(x,y) + S(x+k,y) \ , \tag{3}$$

where A is the albedo of the scene, and S is the actual image the sensor would capture without the projected pattern.

However thanks to the bandpass IR filter, scene information unrelated to the reference pattern R is filtered out so we can assume $S \approx 0$.

By averaging a large amount of captures from different orientations, we assume that all pixels have been exposed to a similar range of albedos, therefore we assume $A \approx 1$.

3.1 Superresolution

Applying the Reverse Stereo Camera Model while using Kinect in its default mode at $VGA@30fps$ is practical as both depth and IR streams are available simultanenously. However we get a low resolution R' with a strong moir pattern.

Capturing several images and fusing them using superresolution partially solves the problem, and individual dots are visible. But the best results are obtained with 1280x1024 IR images and superresolution, although this forces us to toggle between depth and IR modes (Fig. 3).

3.2 Filtering

In most stereo camera systems, gain and exposure settings are synchronized between the cameras, and therefore the illumination from both images is comparable. This is not the case of Kinect, as the reference image does not depend on the actual environment, the brightness is not consistent between views. It is known that this problem is addressed by a prefiltering step, but the actual filter used is not known.

The common way to address this problem is by normalizing both 17x17 patches before matching. We call it the *normalizing* filter.

Using an horizontal Sobel filter is a more computationally efficient alternative [6].

4 Evaluation

It is known that the output from Kinect is very temperature sensitive, and banding artifacts are added by PS1080 to compensate changes in temperature. This effect was noted in [1] and explored specifically in [3]. Although the temperature induced distortions are small, they are enough to corrupt the fine subpixel measurements from this section. Therefore all the experiments were performed when the Kinect temperature was stabilized (1 hour).

On each test case we compare the disparity provided by our model to the disparity provided by Kinect. The mean error, bias and variance are shown. Our results show that the most probable Block Size used is 17x17, as shown in Table 4. On the filter evaluation, the unfiltered system performed the worst, while Normalized and Sobel filters provide similar results (Table 4).

Finally, we measured the impact of the resolution in both pattern and input images. Standard resolution pattern (pS), and infrared (iS) were captured at VGA, while the high resolution versions (pH, iH), were captured at SXGA. The results show that the better results are achieved when the resolution of the pattern and the source image are the same. It is interesting to mention that the system is still usable using only VGA images (mean error well below 1 pixel), this could be of use in those aplications where Kinect must work at 30fps.

	mean error	mean bias	variance
13x13	0.0337	-0.001	0.0047
15x15	0.0219	0.0025	0.0028
17x17	0.0201	0.0066	0.0025
19x19	0.0271	0.0089	0.0037

Table 1. Block Size evaluation. High Resolution Pattern, High Resolution Image, Norm. Filter.

unfiltered	mean error	mean bias	variance
pH, iH	7.6075	7.2096	213.157
pH, iS	31.8239	30.951	484.256
pS, iH	27.0276	25.9541	392.57
pS, iS	16.9155	15.9849	336.521

normalized	mean error	mean bias	variance
pH, iH	0.0201	0.0066	0.0025
pH, iS	3.8322	3.3434	117.028
pS, iH	0.4228	0.3072	11.6447
pS, iS	0.2402	-0.24	0.0054

Sobel	mean error	mean bias	variance
pH, iH	0.0201	0.0059	0.0026
pH, iS	20.34	19.5053	398.292
pS, iH	0.4499	0.3515	13.4228
pS, iS	0.2437	-0.2435	0.0047

Table 2. Comparison of the disparity provided by our model with respect to Kinect. The mean error, bias, and variance are shown when using a Standard or High resolution pattern (pS, pH respectively), and Standard or High resolution IR (iS, iH). Results are shown with three prefiltering steps: None, Block normalization, and Sobel.

4.1 Kinect as a Stereo Camera

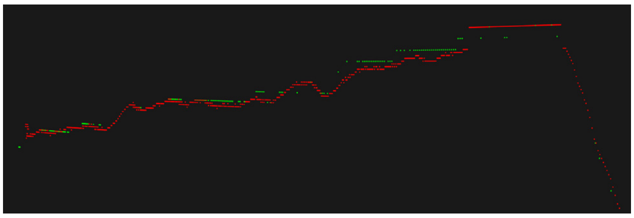

Figure 4. Comparison of Kinect algorithm (green) and SGBM (red). Note the resolution of SGBM.

Any stereo algorithm can be used with Kinect once we have estimated its reference image. This adds versatility to the sensor as it is possible to execute an algorithm that better suits the needs of the scenario.

A basic limitation of Kinect integrated depth map is the fixed point depth resolution, which is limited to 1/8 of a VGA disparity pixel. Other algorithms provide better subpixel precision. We compared Kinect to a Semi-Global Block-Matching [5] by using the reference as left image, and the IR as right, as seen in Fig. 4

5 Conclusions

We have shown a model that can accurately reproduce the depth maps of Kinect. We use this model to estimate the internal reference image that Kinect uses to generate depth. This image allows us to avoid the internal depth pipeline and apply any stereo algorithm. Although this paper focuses on reproducing the Kinect depth output, the results suggest that the pattern projected by the IR laser does not provide an optimal dot density as it has been designed to be inexpensive. It has proven good enough for skeleton tracking in consumer-grade media interaction, but there is ample margin for improvement. Our next step is to formally analyze the limitations of the Kinect model, and develop stereo algorithms to compensate them.

References

[1] M.R. Andersen, T. Jensen, P. Lisouski, A.K. Mortensen, M.K. Hansen, T. Gregersen, and Ahrendt P., *Kinect depth sensor evaluation for computer vision applications*, 2012, Technical Report.

[2] H. Aoki, Y. Takemura, K. Mimura, and M. Nakajima, *Development of non-restrictive sensing system for sleeping person using fiber grating vision sensor*, Micromechatronics and Human Science, 2001.

[3] David Fiedler and Heinrich Müller, *Impact of thermal and environmental conditions on the kinect sensor*, International Workshop on Depth Image Analysis at the 21st International Conference on Pattern Recognition (2012).

[4] D. Herrera C, J. Kannala, and J. Heikkilä, *Accurate and practical calibration of a depth and color camera pair*, Computer Analysis of Images and Patterns, 2011.

[5] H. Hirschmuller, *Accurate and efficient stereo processing by semi-global matching and mutual information*, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005.

[6] H. Hirschmuller and S. Gehrig, *Stereo matching in the presence of sub-pixel calibration errors*, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009.

[7] Kourosh Khoshelham and Sander Oude Elberink, *Accuracy and resolution of kinect depth data for indoor mapping applications*, Sensors (2012).

[8] Kurt Konolige, *Projected texture stereo*, International Conference on Robotics and Automation, 2010.

[9] Kurt Konolige and Patrick Mihelich, *Technical description of kinect calibration*, `www.ros.org/wiki/kinect_calibration/technical`, 2010, [Online; accessed 15-December-2012].

[10] Jae-Han Park, Yong-Deuk Shin, Ji-Hun Bae, and Moon-Hong Baeg, *Spatial uncertainty model for visual features using a kinect sensor*, Sensors (2012).