Microsoft Kinect: features and implementation

Francesco Castaldo [1]

December 23, 2013

[1]F. Castaldo (*francesco.castaldo@unina2.it*) is with the Dipartimento di Ingegneria Industriale e dell'Informazione, Seconda Università degli Studi di Napoli, Aversa (CE), Italy

Contents

Chapter 1

Introduction

Kinect for Xbox360 (formely known as Project Natal) is a motion sensing input device, builds on software tecnology developed by Rare (a subsidiary of Microsoft) and based on range camera tecnology developed by PrimeSense, an Israeli company (Microsoft basically bought this tecnology).

The Kinect sensor is based on the PrimeSensor^{TM} Reference Design and its basic working principles are explained in a patent called "Depth Mapping using Projected Patterns" [1].

Figure 1.1: Kinect for Xbox 360

Chapter 2

Hardware informations

As stated in "The PrimeSensor™ Reference Design[1] 1.08" document, The PrimeSensor™ Reference Design is an end-to-end solution that enables a computer or a system capable of elaboration (TV, ecc) to perceive the world in three-dimensions and to translate these perceptions into a synchronized depth image, in the same way that humans do. The solution includes a sensor component, which observes the scene (the users and their surroundings), and a perception component, or brain, which comprehends the user interaction within these surroundings.

The PrimeSensor Reference Design was created as a low-cost, plug and play, USB-powered device that can either sit on top of or next to a television screen or a monitor, or be integrated into them: Microsoft integrated it in his Kinect peripheral. The Reference Design generates realtime depth, color and audio data of the room scene. It works in all room lighting conditions (whether in complete darkness or in a fully lit room). It does not require the user to wear or hold anything, does not require calibration and does not require computational resources from the host's processor. It is also almost immune to ambient light. PrimeSense's technology for acquiring the depth image is based on Light Coding™. We will see in the following that Light Coding works by coding the scene volume with near-IR light (the IR Light Coding is also invisible to the human eye). The solution utilizes a standard off-the-shelf CMOS image sensor to read the coded light back from the scene. PrimeSense's SoC chip is connected to the CMOS image sensor, and executes a sophisticated parallel computational algorithm to decipher the received light coding and produce a depth image of the scene.

The PrimeSensor Reference Design is built around PrimeSense's PS1080 SoC. The PS1080 controls the IR light source in order to project the scene with an IR Light Coding image. The IR projector is a Class 1 safe light source,

[1] A reference design refers to a technical blueprint of a system that is intended for others to copy. It contains the essential elements of the system; however, third parties may enhance or modify the design as required. The main purpose of reference design is to support companies in development of next generation products using latest technologies.

Figure 2.1: PrimeSensor^{TM} Reference Design

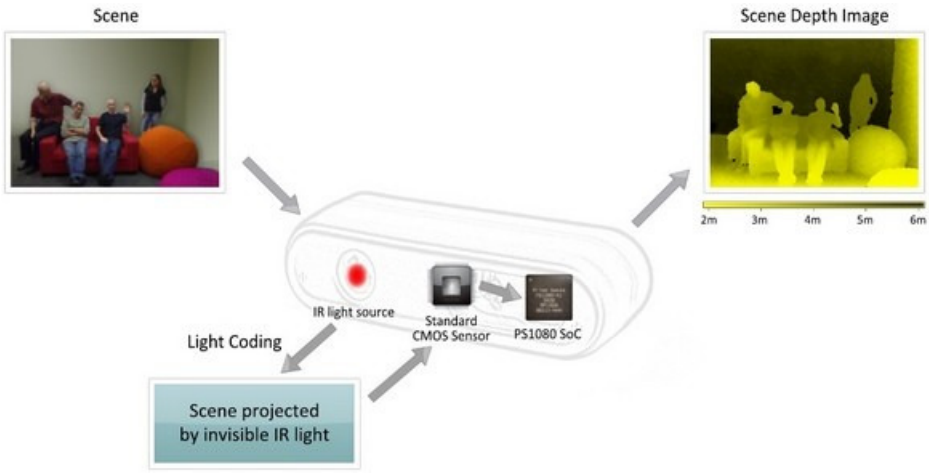

Figure 2.2: How the PrimeSensor works

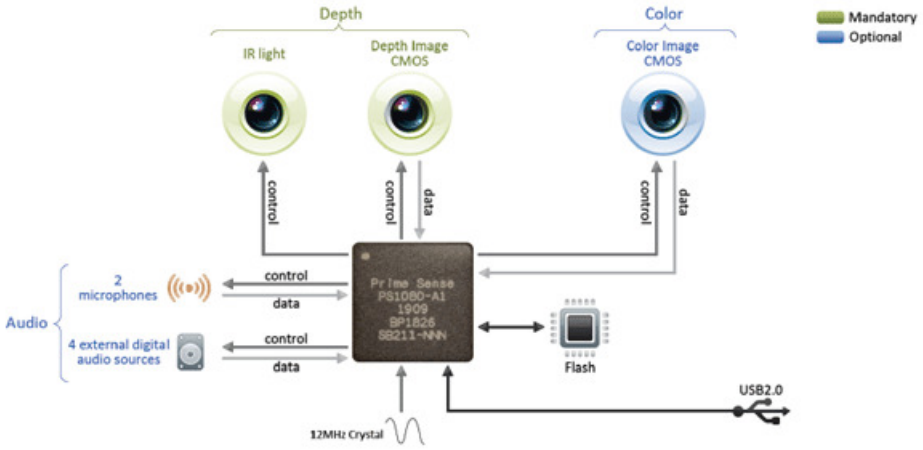

Figure 2.3: PS1080 SoC

and is compliant with the IEC 60825-1 standard. The CMOS image sensor receives the projected IR light and transfers the IR Light Coding image to the PS1080. The PS1080 processes the IR image and produces an accurate per-frame depth image of the scene. The PrimeSensor includes two optional sensory input capabilities: color (RGB) image and audio (the PrimeSensor has two microphones and an interface to four external digital audio sources). To produce more accurate sensory information, the PrimeSensor Reference Design performs a process called Registration. The Registration process's resulting images are pixel-aligned,which means that every pixel in the color image is aligned to a pixel in the depth image. All sensory information (depth image, color image and audio) is transferred to the host via a USB2.0 interface, with complete timing alignment. In summary, The PS1080 SoC receives a Light Coding^{TM} infrared pattern as an input, and produces a VGA-size depth image of the scene as output.

As expected, neither Microsoft nor PrimeSense have released so much material about the device and its internal functioning. OpenKinect [4] is an open community of people interested in making use of Kinect hardware, and they have done some research about the internal characteristic of the peripheral.

The RGB camera is a MT9M112 or a MT9v112. The depth camera is a MT9M001. The Kinect has also a Accelerometer KXSD9, a USB Audio interface TAS1020B and a USB HUB uPD720114.

The illuminator uses an 830nm laser diode. There is no modulation - output level is constant. Output power measured at the illuminator output is around 60mW (Using Coherent Lasercheck). The laser is temperature stabilized with a small peltier element[2] mounted between the illuminator and the aluminium

[2]A peltier element is an electronic device consisting of metal strips between which alternate strips of n-type and p-type semiconductors are connected. Passage of a current causes heat

Property	Spec
Field of View (Horizontal, Vertical, Diagonal)	58H, 45 V, 70 D
Depth image size	VGA (640x480)
Spatial x/y resolution (@ 2m distance from sensor)	3mm
Depth z resolution (@ 2m distance from sensor)	1cm
Maximum image throughput (frame rate)	60fps
Operation range	0.8m-3.5m
Color image size	UXGA (1600x1200)
Audio: built-in microphones	Two mics
Audio: digital inputs	Four inputs
Data interface	USB 2.0
Power Supply	USB 2.0
Power consumption	2.25W
Dimension (Width x Height x Depth)	14cm x 3.5cm x 5cm
Operation environment (every lighting condition)	Indoor
Operating temperature	0 C - 40 C
Weight	750 g

Table 2.1: Kinect specification

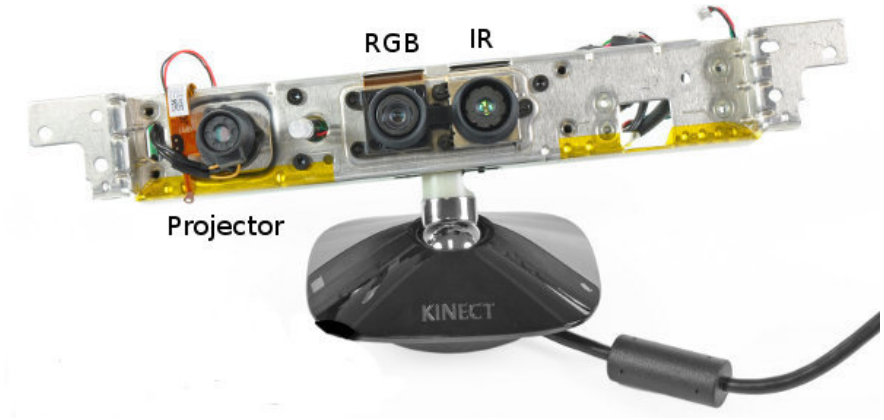

Figure 2.4: The internal structure of the sensor

mounting plate. This element can both heat and cool the laser to maintain a constant temperature, presumably in order to stabilize the laser's output wavelength. on the sample tested, at room temperature, the laser is being slightly heated. The depth sensing appears somewhat sensitive to the relative position of the illuminator and sensor. Even a gentle bend of the aluminium plate causes significant disturbance of the depth image. Loosening the illuminator screws and moving the illuminator slightly has the following effects : rotation gradually narrows the visible width of the frame, down to about 10% - the image appearance is the same, the left & right edges just get progressively cropped to black. Horizontal panning alters just the depth values without changing the overall scene appearance. Vertical panning shifts the visible part (as limited by rotation) of the frame horizontally, again there is no change in appearance of the visible part of the scene. There is a non-resettable 102 C thermal cutout attached to the outside of the illuminator housing - this is connected in series with the main 12V supply to the kinect, so clearly a product-safety feature. This positioning suggests that it may be to detect fault conditions causing the laser to heat its enclosure, e.g. if the front window is damaged, or possibly a peltier-induced meltdown distorting the housing and causing laser light to escape other than through the pattern optics. The raw laser power is definitely eye-hazardous if not spread out by the pattern-generating optics. Apart from the laser diode, the illuminator is likely to contain a temperature sensor and a photo-diode for laser output power feedback - the latter may be integrated in the laser diode can.

The depth sensor uses a monochrome image sensor. Looking at the signals from the sensor, resolution appears to be 1200x960 pixels at a frame-rate of 30Hz. The camera's I2C control interface performs a one byte transaction every frame - this could be something like gain setting or average level sensing. No change in this data was seen during random hand-waving in front of the sensor. The camera has an IR-pass filter at the laser wavelength - tests with various light sources show minimal sensitivity to visible and 950nm sources.

Marvell chip is for audio processing, and has nothing to do with the depth system. Markings on chip are 88AP1-BJD2 P2G2750A/2 1024 AOP CX08 88AP102. Possibly one of Marvell's ARMADA series ARM chips, which has members with clock rates from 400MHz to 1GHz. External clock is 26MHz. Connected to the chip are a Winbond 25Q16B 16mbit, quad SPI flash, a custom-marked 8 pin device marked H102338 XBOX1001 X851716-006 GEPP , and a 512mbit DDR2 SDRAM. The SPI flash contains the firmware, which apperas to resemble 32 bit ARM code. The Marvell is also responsible for cryptographic authentication of the Kinect to the Xbox (to prevent clones). It would make sense for the custom 8-pin device to be the authentication chip containing the RSA key and certificate.

The cooling fan only comes on when the internal temperature exceeds about 70 C, measured by thermistor RT3 near the fan connector on the camera board. Once on it stays on for at least a few tens of minutes (maybe until power-down).

to be absorbed from one set of metallic strips and emitted from the other by the Peltier effect

The fan is powered by the 5V supply from the USB connector.

All the image and depth sensor functionality is on the front PCB, and it should be possible to operate this without the rear board, using an external power supply to provide the +5V, 3.3V and +1.8V rails. It has been confirmed that connecting the USB pins directly to a PC USB port works, with only the power connections from the rear PCB remaining connected. Unfortunately the peak current draw puts it slightly above what can be 'legally' drawn from a single USB connector, due to the current draw of the Peltier element. Total power = 10mA @5V, 200mA @1.8v, 900mA @3.3V, = 3.4W, which would pull about 750mA at 5V assuming 90% converter efficiency and a 5V supply from the USB port. In practice cable voltage drops on the USB supply would mean near 800-850mA. It may be feasible to limit this, e.g. by adding a series resistor to the peltier to limit its maximum power draw. It may be possible to operate without the peltier - this is something still to be investigated.

Connector on PCB appears to be Tyco FP series Mating connector Tyco 4-174639-4 It appears that part 4-175638-4 is the same connector supplied on tape/reel, with a lower min order qty of 1500 vs. 10K for the 4-174639-4 part Pins 1,3,5 : +3.3V 500-900mA depending on Peltier temperature Pins 2,4,6,9,12 : Ground Pin 7 : +1.8v, 200mA Pin 8 : USB DP Pin 10 : USB DM Pin 11 : +5V 10mA directly from USB connector. The fan, when on, is also powered by this supply. Pin 13 : Input to sensor board, pulled weakly to +3.3v. Rear board pulls low until USB connection established. Can be left unconnected. Pin 14 : Output from sensor board - pulses high for about 70ms after pin 13 rising edge. Can be left unconnected.

Chapter 3

Working principles

As mentioned above, The PrimeSensor solutions consist of a chip (PrimeSense PS1080 SoC), a 3D depth sensors (IR light source and CMOS image sensor) and a RGB camera (color image sensor); it is also present an audio system (two microphones), although we are not going to use it. The PS1080 SoC acquires the depth image by directing invisible infrared light at the objects; the CMOS sensor then reads the coded light back from the scene. But in which way the Kinect reconstruct the depth image? To recover the information of depth, we normally need a stereo vision system (two cameras), but the Kinect has only a IR light source, a Depth camera and a RGB camera.

3.1 3D Scanner and Stereo Vision System

Laser range finders, time-of-flight laser scanning, stereo vision cameras are all systems that resolve, in different ways and with different results, the problem of depth perception, a fundamental topic of the Computer Vision. A first distinction is between 3D scanners (range finders, structured-light 3D scanners, triangulation laser scanner etc) and cameras (at least two). 3D scanners are very analogous to cameras: like cameras, they have a cone-like field of view and they can only collect information about surfaces that are not obscured; the difference is that, while a camera collects color information about surfaces within its field of view, a 3D scanner collects information about surfaces in its field of view. The purpose of a 3D scanner or a stereo vision camera is to create a point cloud of geometric samples on the surfaces of the objects; we obtain it from a depth map (or range image), that is an image whose pixel values represent a distance or depth from the sensor's origin. Range imaging is the name for a collection of techniques which are used to produce a 2D image showing the distance to points in a scene from a specific point, and all the devices presented above (3D scanners, stereo cameras) can be seen as different type of instruments that produce range image (range cameras). With knowledge of the camera's intrinsic calibration parameters, a range image can be converted into a point cloud. The

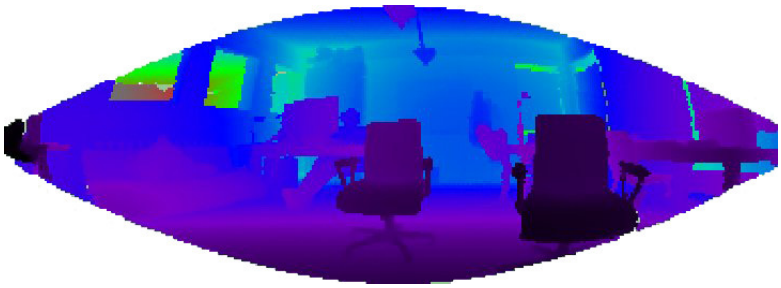

Figure 3.1: A range image (or depth map)

points of the point cloud can be used to extrapolate the shape of the objects (in a process called reconstruction).

A system can be active or passive. Active vision refers to techniques that use a controlled source of structured energy emission, such as scanning laser source or projected pattern of light, and a detector, such as a camera. A common example of active vision is laser range finding, where an active source moves in a environment in order to scan the surface of objects in the scene. On the other hand, passive vision refers to techniques that not use a specific structured source of energy in order to form an image; no energy is emitted for the purpose of sensing, it is only received. Both this approaches have good features but also some weaknesses: inaccurate results of stereo imaging caused by the ambiguity problem in passive vision, slow, noisy and sparse results of laser scanners in active vision. A basic principle in both active or passive vision devices is the triangulation principle. In active vision, a triangle is created between the light, the object and the sensor (e.g. a CCD camera). In passive vision, a triangle may be created between the object and two sensors (e.g. cameras).

The Kinect, we will see it later, is, for certain aspects, a 3D scanners (a structured-light 3D laser scanners): but for the reconstruction of the scene it use concepts and tools of stereo vision system, in particular the creation of a disparity map.

3.2 Functioning of the Kinect

The working principles of the algorithm of reconstruction of the scene performed by the Kinect are explained in the 2008 PrimeSense patent "Depth Mapping using Projected Patterns"[1], that also redirects to another Primesense patent (of 2007), named "Method and System for Object Reconstruction"[2]. Other information are available on the ROS (Robot Operating System) website[3].

In the 2008 patent is described an apparatus for mapping an object includes an illumination assembly, which includes a single transparency containing a fixed pattern of spots. A light source trans-illuminates the transparency with optical radiation (IR light source), so as to project the pattern onto the object. An image capture assembly recovers an image of the projected pattern onto the

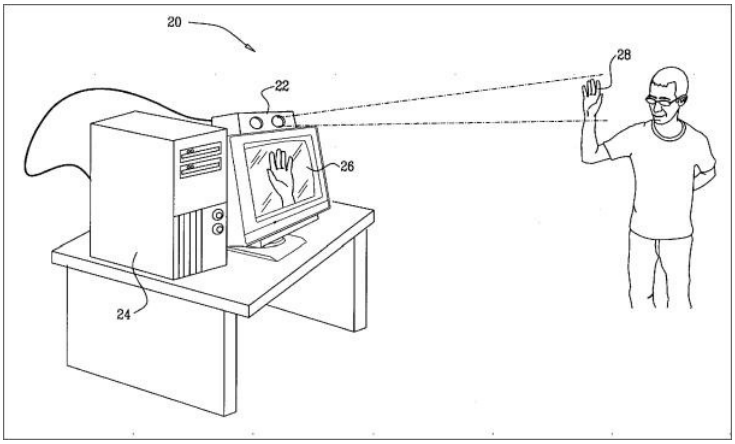

Figure 3.2: Depth mapping using projected patterns

object. A processor processes the captured image so as to reconstruct a 3D map of the object. So the main idea is to project this huge point pattern on the framed scene, and then recover the map of the environment by measuring the distortion on the pattern. This is possible because the projected pattern, which at first sight seems to be randomly distributed (a pseudorandom pattern) , is known by the kinect; we can see it pointing the Kinect towards a flat surface (e.g. a wall), turning off the lights and framing the area with a digital camera. As we can see in Figure 3.7, the pattern is formed of light and dark speckles, and it is generated from a set of diffraction gratings, with special care to lessen the effect of zero-order propagation of a center bright dot. There are also nine bigger points, distributed along the pattern, that should facilitate the subsequent procedure of localization of each points of the projected pattern.

On the ROS website, there are consideration about the retrieval of the depth. Depth is calculated by triangulation against this know pattern (memorized at a know depth). For each pixel in the image (IR image), the Kinect use a small correlation window (9x9 or 9x7) in order to resolve the correspondence problem and compare the local pattern at that pixel with the memorized pattern at that pixel: the bes match gives an offset from the know depth, and this is the disparity for the Kinect. Given the known depth of the memorized plane and the disparity, the estimated depth for each pixel can be calculated by triangulation, as shown above. However, we should note that the Kinect raw disparity is not normalized as we have seen previously. In fact we know that, for a stereo system in which the cameras are calibrated so that the rectified images are parallel and have corresponding horizontal lines, the relationship between disparity and depth is:

$$Z = \frac{b \cdot f}{d}$$

At zero disparity, the rays from each camera are parallel, and depth is infi-

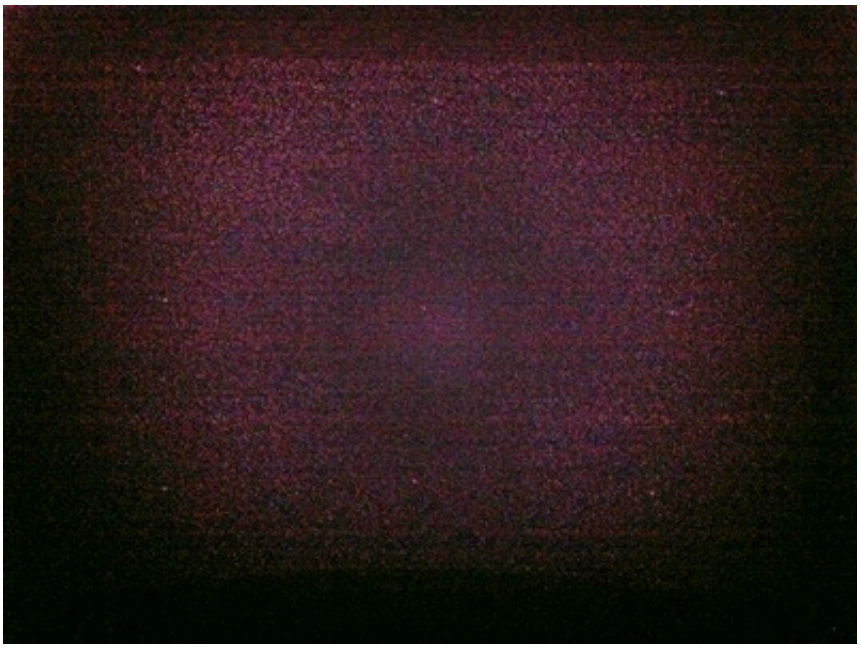

Figure 3.3: The projected patterns by the Kinect

nite. Lager values for the disparity means shorter distances. The Kinect, on the contrary, returns a raw disparity in which a zero value does not correspond to infinite distance: the Kinect disparity and the normalized disparity are related by this equation:

$$d = 1/8 \cdot (d_{off} - k_d)$$

where d is a normalized disparity, k_d is the Kinect disparity, d_{off} is an offset value particular to a given Kinect device; the $1/8$ factor appears because the values of k_d are in $1/8$ unit pixel. A monocular calibration of the IR camera (we are only considering the depth camera and not the RGB camera) finds the focal lenght, the distortion parameters and the lens center of the camera; from this information, it is possible to calculate the baseline and the d_{off}. Tipical value of this quantities are:

- $b \sim 7.5\, cm$
- $d_{off} \sim 1090$
- $f \sim 570\, pixel$

Other information are included in the 2007 patent, that presents a generic system (we are far away from the Kinect release) but explains also in general the algorithm of the scene (in the article they refer to a single object) reconstruction. The system comprises an illuminating unit, and an imaging unit.

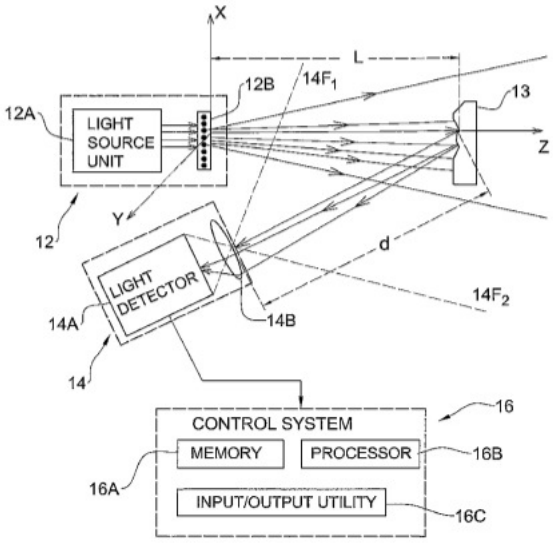

Figure 3.4: Method and System for Object Reconstruction

The illuminating unit comprises a coherent light source and a generator of a random speckle pattern[1] accommodated in the optical path of ilumuminating light propagating from the light source towards an object, thereby projecting onto the object a coherent random speckle pattern. The imaging unit is configured for detecting a light response of an illuminated region and generating image data. The image data is indicative of the object with the projected speckles pattern and so indicative of a shift of the pattern in the image of the object relative to a reference image of said pattern. This enables real-time reconstruction of a three-dimensional map of the object.

In the first part of the article there is a useful explanation about the object reconstruction techniques developed in this field. One of the approaches deals with triangulation (we already explained it) using two cameras observing the same object: a relative shift of the same items in the images acquired by the cameras is related to a distance to these items. Another known technique of the kind specified utilizes numerical algorithms based on the use of shadows of edges in the single captured image in order to compute the 3-D map of an object. Another approach is based on projection of patterns: some of these techniques utilize projection of a line onto an object and scanning the object with this line. A curvature generated in the line image is indicative of the 3-D map of the object. Yet another techniques, based on the projection of patterns, include single projection of a 2-D periodic pattern; in this case, 3-D details of the object shift the lines of the periodic pattern in the captured image. A relative shift

[1] A speckle pattern is a random intensity pattern produced by the mutual interference of a set of wavefronts

of these lines is related to the 3-D information of the object. All this methods suffer of several drawbacks: the triangulation has low 3-D resolution and does not provide real-time mapping because classification and registration are high-level processing operations; the numeric algorithms accumulate errors and also require high-level processing; the projection of line patterns does not provide real time reconstruction (it takes time to scan the object with the line) and also the estimation becomes more distorted in case the object moves; finally, the projection of the 2-D periodic pattern suffers from the fact that movements larger than the period of the projected pattern cannot be distinguished, and also there is a defocussing of the pattern at large distance.

After this introduction, there is the explanation of the invention, defined as a technique that allows a real-time and very accurate mapping of 3-D objects, which can be achieved with a very simple and inexpensive optical set up. In the article "object reconstruction" means the acquisition of 3-D information of any part or whole of the object surfaces, and "real-time" refers to an operation for which the combined reaction-operation time of a task is shorter than the maximum delay that is allowed. The main idea of the invention consists of utilizing projection of a laser random speckle pattern onto an object the 3D surface data of which is to be reconstructed. A speckle patter is a field-intensity pattern produced by the mutual local interference of partially coherent beams; the brighter spots are positioned where light was scattered in phase, and the dark position are positioned where light was in anti phase. Laser speckles are random self-generated patterns, and this pattern is "constant", i.e. the pattern is substantially not varying along the Z-axis within the region in which the 3D measurements are to be taken. We can also distinguish between "primary speckles" (the speckles projected onto the object) and "secondary speckles", which are typically associated with surface roughness and aperture of the imagins lens. Another interesting properties inherent to the laser speckle is that its projection on an object yields highly contrast images. This is because a laser speckle pattern is create with high contrast, and high contrast images can be represented taking 0 or 1 values for each pixel (binary image); this high contrast allows for reduction of processed data and faster image reconstruction. As we already said, the 3-D map of the object is estimated by examining the relative shift of a laser random pattern relative to a reference image of said pattern.

The optical set up can be very simple and cheap: it can include only a small coherent light source (laser) and a pattern generator in the form of a light diffuser. The light source may be constituted by a light emitting assembly (laser) or a light guiding arrangement (optical fiber) associated with a remote light emitting assembly. The pattern generator is a light diffuser, for example a ground glass. The light detector is a pixel matrix, e.g. a CCD, with an imaging lens arrangement. The control system is typically a computer system having a memory utility, a data processing and analyzing unit and a I/O unit (such a display). The control unit is configured for storing the reference data indicative of a reference image of the speckle pattern. The reference data is indicative of the image of the speckle pattern acquired at a reference plane oriented normally to the optical path of illuminating light propagation and at a

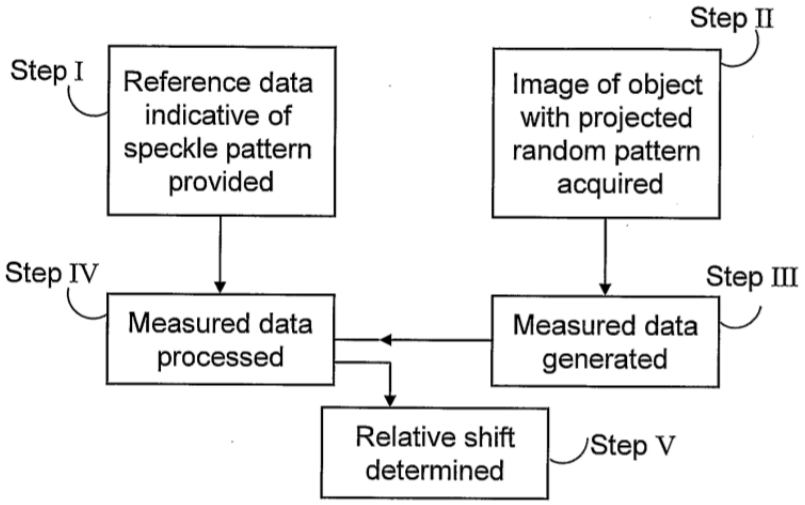

Figure 3.5: Steps of reconstruction process

substantially the same distance from the diffuser and imaging unit. The control unit is programmed for processing and analyzing the image data utilizing the reference data for determining correlation between the object and the reference images. An important parameter is the average speckle feature size, Δx_{cam}, on a pixel plane of the imaging unit is at least two pixel: if the diffuser and imaging unit are oriented so as to provide substantially equal distance from this two units to the object, we can determine Δx_{cam} as:

$$\Delta x_{cam} = \frac{F}{\phi_D}\lambda$$

, with F the focal lenght, ϕ_D the illuminating spot size on the diffuser, and λ the wavelenght of the illuminating light. This formula is valid only if the distance between the diffuser and the object (L) and the distance between the camera and the object (d) are the same; otherwise, the formula is:

$$\Delta x_{cam} = \frac{\lambda F}{\phi_D} \cdot \frac{L}{d}$$

In the next figure are shown the main steps of the reconstruction method.

In Step I the reference data indicative of the image of a speckle pattern (image of the pattern with no object) is provided ans stored in the memory utility of control system. In order to do this, the speckle pattern is projected onto a certain region (where the object will be placed) and the reference image of the speckle pattern is acquired. In Step II the image of the object is acquired, and in Step III the measured data is generated. To this end, the object is

illuminated by the light carrying the speckle pattern, and a light response from the object is collected by the image lenses and detected by the light detector. In Step IV there is the confront between the measured and the reference data, in order to determine a relative shift of the features of the random pattern in the object image, relative to the pattern in the reference image (in Step V). We can achieve this using a matching algorithm, e.g. a correlation algorithm, with a small moving window scanning the captured image and correlating it to the reference image. The correlation peak indicates the relative shift, from which we can extract the 3-D information.

The last part of the article shows examples about the algorithmic constellation suitable to be used in the data processing; they are required to have low computational complexity and they have to allow real time 3-D mapping. We already said that, in order to obtain one specific depth point, we have to calculate the transversal shift of its vicinity from the reference pattern. One possible implementation of this matching is by searching the correlation peak, e.g. taking one or more windows from the image (the neighborhood of the inspection point) and computing a match of the window with the reference image. The windows for the correlation can be either constant (e.g. 16x16) or modified in accordance with the local properties. Afterwards we can use different algorithms to execute the actual 3-D reconstruction: this various approaches differ from each other in complexity, accuracy etc. Among the algorithms, in the article is presented a prediction-based region-growing method, that offers a good tradeoff between complexity and performance. It is based on the consideration that two close points on one object are usually characterized by a small height (along Z axis) difference between them (the "object continuity assumption"). Therefore we can predict the depth value of a point from its neighboring points on the same region (object). This prediction is of course tested and refined, and if it is found to be adequate, the point is joined to the given region; otherwise, the point under ispection belongs to a different region from its neighboring points.

The algorithm presented before contains the following steps:

1. Relevance/shadow preparation step: pick a sampling point on the picture (the image of the object) and determine whether the sampling point is a SHADOW point or a relevant speckle window to be marked as UNKNOWN.

2. Obtain a new region anchor step: while the number of UNKNOWN points is more than a predefined threshold (percentage of the total output points), we choose a random UNKNOWN point and execute a matching to the reference image (by correlating the window around the chosen point with the reference image). Simultaneously, we search for a point with a value of normalized correlation higher than a certain threshold, and if it does, it is assumed that this is the shift of the speckle; this point is called region anchor, and the region growth is attempted around it. If it doesn't, we have to do further optional investigation.

3. Region growing step: a FIFO plurality of ACTIVE points is used, where

each of this points (active points) has already been correlated with the reference and the correlation proved to be successful. Each step one point from the FIFO is fetched and its four neighbors (left/right/up/down) are checked. If one of this neighbors is UNKNOWN, we correlate the windows with the reference, and if the results are good, the point is marked as ACTIVE and added to the FIFO set; if not, the point is marked as EDGE. When the FIFO is emptied, the region is created and the process return to step 2, to grow a new region.

4. Region competition: a optional step in which we can improve quality of region edges (an example is to attempt growing a region not only in UNKNOWN point, but to any point belonging to a different region and so with depth discontinuity).

In conclusion, in the light of what we have seen, it is not easy to categorize the Kinect sensor in one of the categories we have previously discuss. Probably the best choice is to categorize it as a structured-light 3D laser scanner, because it project a pattern of dots (a pseudo-random pattern) on the objects, and construct the depth map by measuring the difference of this pattern from its memorized pattern (at a well know depth); the way in which this task is accomplished is using a particular triangulation technique, in which the confront is made between the pixel of the framed image and the pixel of a image memorized in the device itself.

Bibliography

[1] Depth Mapping using Projected Patterns, Pub. No.: US2008/0240502 A1, Pub. Date: Oct. 2,2008

[2] Method and System for Object Reconstruction, Pub. No.: WO2007/043036 A1 Pub. Date: Apr. 1,2007

[3] ROS website http://www.ros.org/wiki/kinect_calibration/technical

[4] OpenKinect website http://openkinect.org