A Time-Of-Flight Depth Sensor – System Description, Issues and Solutions

S. Burak Gokturk, Hakan Yalcin, Cyrus Bamji
Canesta Inc.
{bgokturk,hyalcin,cbamji}@canesta.com

Abstract

This paper describes a CMOS-based time-of-flight depth sensor and presents some experimental data while addressing various issues arising from its use. Our system is a single-chip solution based on a special CMOS pixel structure that can extract phase information from the received light pulses. The sensor chip integrates a 64x64 pixel array with a high-speed clock generator and ADC. A unique advantage of the chip is that it can be manufactured with an ordinary CMOS process. Compared with other types of depth sensors reported in the literature, our solution offers significant advantages, including superior accuracy, high frame rate, cost effectiveness and a drastic reduction in processing required to construct the depth maps. We explain the factors that determine the resolution of our system, discuss various problems that a time-of-flight depth sensor might face, and propose practical solutions.

1. Introduction

Computer vision field regularly deals with various applications such as tracking, recognition, image understanding, etc. One direction of research has been towards depth sensing as opposed to using regular image intensity cameras, because depth information makes the aforementioned applications more feasible and robust. This paper presents a novel system for depth sensing based on time of flight (TOF) which we believe is the foundation for a new *electronic perception techology*, giving eletronic devices of all types the ability to perceive and interact with the world around them.

TOF systems have been used in radar and Lidar applications for more than thirty years; the reader is referred to [11] for a review. The basic principle involves sending out a signal and measuring a property of the returned signal from a target. The measured property is used to determine the time of flight, and the distance is obtained via multiplication of the time of flight and the velocity of the signal in the application medium.

Our TOF-based depth sensor uses light as its signal, and measures the phase shift of a modulation envelope of the light source as its property. The distance to objects in the scene can be calculated using the properties of light and the phase shift. The depth sensor is implemented in a single chip using an ordinary CMOS process. Its depth resolution is in the order of a few millimeters, and does not require any computationally complex post-processing.

Like any practical method, there are issues during the operation of such a system. For instance, the noise behavior is dependent on the amount of light reflected into the sensor. The resolution of the sensor is improved as more noise is eliminated. There are various techniques to reduce the amount of noise. These are well-known techniques in radar sciences, and mostly depend on spatial and temporal averaging techniques.

An important issue for TOF sensors is the aliasing effect arising from the periodicity of the modulated signal whereby the distances to objects differing in phase by 360 degrees of phase shift are not distinguishable. Using multiple frequencies is a common way of dealiasing such data, and is discussed in the paper. Other issues covered in the paper include improvements to the sensor's dynamic range and methods for eliminating artifacts due to motion and ambient light.

The paper is organized as follows. First, we review previous work on other types of depth sensors. Next, we describe our depth sensor followed by a theoretical analysis of resolution and metrics that affect the resolution. Next, we describe various issues in the operation of such a system, and we provide practical solutions. We then present our experiments and a discussion where we compare our sensor to other types of depth sensors.

1.1 Previous Work on Depth Sensors

There are various camera-based techniques in the literature to measure range. These include triangulation systems such as stereo-vision, (or structured-light), depth-from-focus, depth-from-shape, and depth-from-motion systems. Each of these systems has advantages and disadvantages as described next.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

Triangulation systems measure the distance to objects by analyzing the triangles constructed by the projection rays of two optical systems. Given a point on a visible surface in the world, two optical systems determine the angles α_1 and α_2 formed by the projection rays that connect the surface point with the centers of projection of the two optical systems (Figure 1). Together with the baseline, these two angles determine the shape of the triangle completely, and simple trigonometry yields the distance to the surface point. A major disadvantage of triangulation systems is the necessary baseline to operate. This induces a minimum size limitation on a triangulation system.

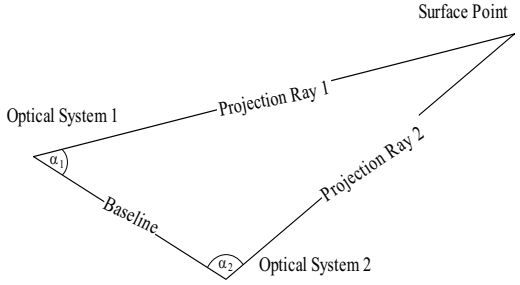


Figure 1. Triangulation determines depth by solving for the height or a side-length of a triangle.

There are two main classes of triangulation systems: passive and active. In passive systems, two cameras are used (hence the name stereo) [1]. These systems require solving the so-called correspondence problem, which amounts to determining which pairs of points in two images are projections of the same point in the world. This is a very complex problem and the solution is computationally expensive. In addition, due to the geometry of triangulation systems, the resolution drops drastically as objects move away from the camera.

Active systems employ one camera and one structured light emitter [2,3]. The structured light system may be any form of light with known pattern. In order to apply triangulation, the projected light pattern needs to be well differentiated from the other objects and ambient light falling on the scene. This requires that the projected light be high powered and well focused. In many cases, it also requires scanning the light through the scene, which makes it difficult to obtain high frame rates.

Another direction of research for obtaining depth information is through depth-from-X methods. In depth from focus methods [4,5], depth is determined by varying the focus of the camera. The frame rate might be limited since multiple images with different camera focus parameters might need to be obtained.

The depth-from-shape method requires prior knowledge of shape. It infers the depth through the appearance of shapes in the image. This method is weak due to its inherent assumption about the knowledge of shapes. Another alternative is to iterate through shape and depth by fixing one and solving for the other in each iteration. This method, unfortunately, gets into ambiguities and singularities if the underlying shape is different from the assumed shape.

The depth-from-motion method calculates depth by measuring the motion of objects [6]. Similar to the depth-from-shape method, this is a weak assumption, and iterative methods often result in singularities.

1.2 Time-of-Flight Depth Sensors

A typical TOF sensor consists of a modulated light source such as a laser or LED, an array of pixels, each capable of detecting the phase of the incoming light, and an ordinary optical system for focusing the light onto the sensor (Figure 2). The light is given a modulation envelope by rapidly turning the light source on and off. Distance measurement is achieved by measuring the phase of the modulation envelop of the transmitted light as received at the pixel array. Although square waves are utilized in practice for modulation, here we use sinusoidal waves for ease of explanation.

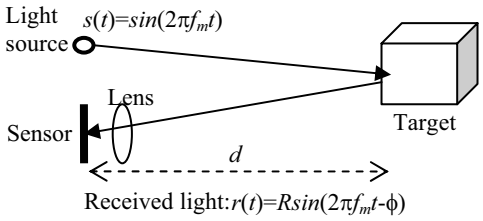


Figure 2. Time of flight measurement.

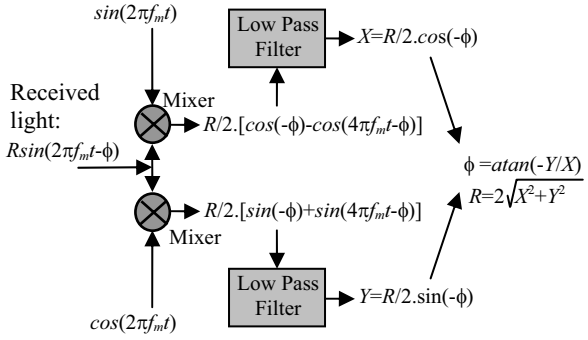


Figure 3. A method of phase/amplitude calculation.

Let $s(t)=sin(2\pi f_m t)$ be the transmitted light where f_m is the modulation frequency. The light reflected from the target falls on a sensor pixel with a phase shift ϕ:

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $20.00 © 2004 IEEE
0-7695-2158-4/04 $20.00 (C) 2004 IEEE

$$r(t) = R\sin(2\pi f_m t - \phi) = R\sin(2\pi f_m (t - \frac{2d}{c}))$$

where R is the amplitude of the reflected light, d is the distance between the sensor and the target, and c is the speed of light, $3x10^8$ m/s. The distance d can be calculated from the phase shift as follows:

$$d = \frac{c\phi}{4\pi f_m}$$

The maximum unambiguous phase delay that can be detected using TOF is a full cycle of the modulation period, which corresponds to an unambiguous range of $c/2f_m$. For example, the maximum unambiguous range for f_m=50 MHz is 3m.

The phase and the amplitude of the reflected light can be extracted via signal processing techniques such as the one given in Figure 3. However, the method of Figure 3 requires mixers and low-pass filters, which can only be implemented using complicated circuitry. In practice, phase detection can be implemented more efficiently as described in [7,8,9,10,11].

Various TOF sensors have been reported in the literature. The system of [7] employs a CCD sensor and reportedly achieves 10-cm resolution using a modulation frequency of 15Mhz. The camera system made by 3DV Systems [8] also uses a CCD camera but is coupled with an external shutter. Both of these systems are CCD-based, which is a major obstacle to building single chip, cost effective, and widely available TOF sensors.

2. CMOS Sensor Chip

We have developed a special sensor that takes advantage of device-level charge processing to efficiently implement TOF. The sensor has a 64x64 pixel array and is implemented on a single chip using ordinary, low-cost CMOS process. The sensor chip also incorporates an ADC and circuitry to generate the high-speed modulation signals. The architecture of the chip is shown in Figure 4. The sensor achieves frame rates of up to 50fps and a depth resolution of a few milimeters.

The key part of the sensor design is the special pixel structure. A cross-section of the pixel is shown in Figure 5. The differential structure accumulates photo-generated charges in two collection nodes using two modulated gates. The gate modulation signals are synchronized with the light source, and hence depending on the phase of incoming light, one node collects more charges than the other. At the end of integration, the voltage difference between the two nodes is read out as a measure of the phase of the reflected light. Thus, in effect, this pixel simultaneously performs the function of mixing and low pass filtering. The reader is referred to [9] for details on pixel operation.

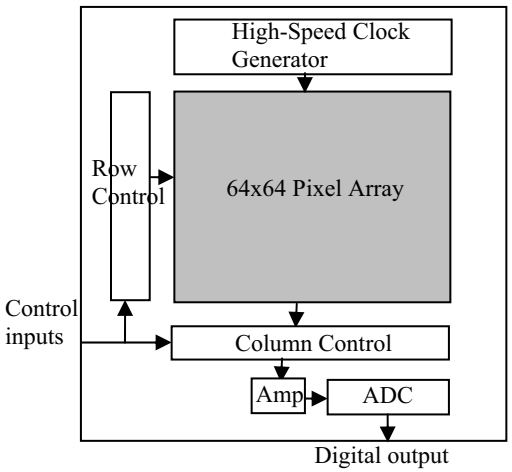


Figure 4. CMOS sensor chip architecture.

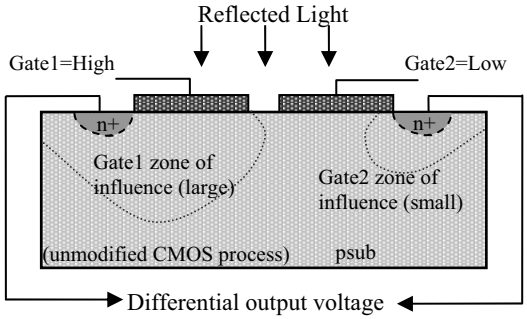


Figure 5. Cross section of a pixel.

In order to reliably detect and disambiguate the phase and amplitude of the reflected light, the light pulses are sent out with two different phase shifts. In other words, measurements are performed two times in one frame, one with a phase shift of 0 degrees and another with 90 degrees for transmitted light. With data obtained using 0 and 90 degree phase shifts, we can determine the phase delay of light with little or no dependence on many factors including object reflectivity, absolute amount of reflected light power, shutter time, and moderate amounts of ambient light. Let V^0 and V^{90} be the pixel values obtained with 0 and 90-degree phase shifts respectively. Phase calculation is done using the inverse tangent function:

$$phase = \arctan(V^{90} / V^0)$$

A linear relationship exists between the phase values and actual distance, as shown in Figure 6. The sensor is calibrated to accurately characterize the linear relationship on a pixel-by-pixel basis. Calibration is done by collecting data using a planar target that is gradually moved within a full range of the modulation

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04$20.00 (C) 2004 IEEE
0-7695-2158-4/04$20.00 (C) 2004 IEEE

frequency. The distance and resolution results reported in this paper are obtained with a calibrated system.

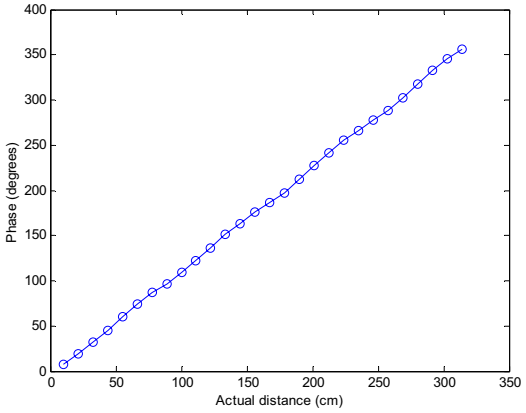


Figure 6. Phase-distance relationship for the depth sensor.

3. Analysis of Resolution

An important attribute of a depth sensor is its depth resolution. This section analyzes the factors that affect the resolution of our system. For simplicity of the analysis, we assume that the pixel output is single-ended voltage even though it is actually differential.

Let P_{laser} be the optical power of the light source, and A the total (target) area illuminated. The amount of laser light received by the sensor depends on the reflectivity of the objects in the direction of the sensor. This reflectivity is denoted by r. The total number of electrons generated in one pixel at the end of an integration (shutter) period of T can be written as

$$N_{electrons} = \frac{P_{laser} k_{opt} q_e r T}{A}$$

where q_e is the quantum efficiency and k_{opt} is a constant determined by the properties of the optical system including lenses, diffuser, pixel size, etc. The electrons collected by the pixel are stored in a storage capacitor C, whose voltage at the end of the integration period is the signal produced by the pixel as a measure of phase delay (distance). Due to the interaction of the reflected light with the reference (modulation) signal in the pixel as described in Section 1.2, only a fraction p of the electrons contribute to this signal. Essentially p represents the *phase overlap* between the reflected light and the reference signal. Hence the voltage across the storage capacitor at the end of integration is

$$V_{signal} = \frac{qpN_{electrons}}{C} = \frac{qpP_{laser} k_{opt} q_e r T}{CA}$$

where q is the charge of an electron, 1.6×10^{-19} C. The value of p at a particular pixel depends on the distance between this pixel and the target it is imaging. Depending on the distance, and hence p, V_{signal} changes from $V_{signal}(p=0)$ to $V_{signal}(p=1)$. Since $V_{signal}(p=0)=0$, the total voltage swing V_{swing} is equal to $V_{signal}(p=1)$.

One of the components of noise on V_{signal} is the shot noise. There are other sources of random or pseudo random noise as well. These include ADC quantization noise, kT/C reset noise, thermal noise etc. In our system shot noise is usually the dominant source of error and determines the accuracy of depth measurements. The error in V_{signal} due to shot noise is calculated by projecting the uncertainty in the number of electrons to the voltage across the storage capacitor. In turn this results in an RMS (root mean square) voltage error of

$$V_{noise} = \frac{q\sqrt{pN_{electrons}}}{C} = \frac{q}{C}\sqrt{\frac{pP_{laser} k_{opt} q_e r T}{A}}$$

The voltage resolution of the sensor is calculated by V_{swing}/V_{noise}, which depends on p in V_{noise}. For resolution analysis, we use $V_{noise}(p=1)$ since it maximizes the magnitude of noise. Thus, from the voltage resolution, the depth resolution is determined by the number of small divisions that the unambiguous range can be reliably divided:

$$resolution = \frac{Range}{V_{swing}/V_{noise}(p=1)} = \frac{c}{2f_m}\sqrt{\frac{A}{P_{laser} k_{opt} q_e r T}}$$

The finite resolution implies that the distance value measured by each pixel deviates from the correct value by an amount whose statistical standard deviation is given by the resolution equation. Changing each parameter in the resolution equation involves a tradeoff. For instance, a high power laser results in better resolution, but increases the electrical power consumption and the cost of the system. Resolution can also be improved by reducing the imaged area, or by increasing the target reflectivity r, or by increasing the integration time, T. The integration time cannot be increased arbitrarily, however; since it determines the frame rate for which most applications have a minimum requirement. It is also possible to improve resolution with q_e by using a lower-wavelength laser. However, keeping the wavelength in the infrared range is desired by most applications.

Another way of improving resolution is to increase the modulation frequency f_m; although this reduces the unambiguous range, resulting in a higher degree of aliasing. See Section 4.2 for a discussion of aliasing.

In the presence of ambient light the above resolution equation needs to be revised since ambient light contributes to the (shot) noise, but not to the useful signal. Let P_{amb} be the ambient light power present in the target area A. Factoring in P_{amb}, we find that the RMS error in the voltage across the storage capacitor becomes

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04$20.00 (C) 2004 IEEE
0-7695-2158-4/04$20.00 (C) 2004 IEEE

$$V_{noise-amb} = \frac{q}{C}\sqrt{\frac{p(P_{laser}+P_{amb})k_{opt}q_e rT}{A}}$$

Since the useful signal V_{signal} remains the same, the depth resolution changes to

$$resolution = \frac{c}{2f_m}\sqrt{\frac{P_{laser}+P_{amb}}{P_{laser}^2}\frac{A}{k_{opt}q_e rT}}$$

Clearly, to get the best resolution we must minimize P_{amb}. Thus for the sensor to operate in the presence of sunlight some techniques have to be developed. This issue is further discussed in Section 4.4.

4. Issues

Below we describe the issues that may affect a TOF system such as ours and propose various solutions.

4.1. Improving Noise Behavior

The resolution analysis given above is a per-pixel, per-frame resolution. The effect of limited resolution is seen as a noisy behavior on the depth images, where the standard deviation of the noise is equivalent to resolution of the system. The noise can be reduced using techniques from radar sensing, in particular through spatial and temporal averaging.

The output range of each pixel can be modeled as a Gaussian with a mean value of μ_i and standard deviation of σ_i where μ_i corresponds to the actual range value of pixel *i*. Assume that the range value is constant over a local neighborhood of pixel *i*, and that all pixels have the same noise behavior modeled by a Gaussian $(\mu_k, \sigma_k)=(\mu, \sigma)$. Let the range value be obtained via averaging of a neighborhood (spatial or temporal) of pixels around pixel *i*. According to probability theory:

$$Mean\left(X = \frac{1}{N}\sum_{k=1}^{N} X_k\right) = \frac{1}{N}\sum_k \mu_k = \mu_i$$

$$Std(X) = \frac{1}{N} Std(\sum_{k=1}^{N} X_k) = \frac{1}{N}\sqrt{N}\sigma_i = \frac{\sigma_i}{\sqrt{N}}$$

In other words, the resolution can be increased by a factor of $\sqrt{N}$ if the range values are averaged in a spatial or temporal neighborhood of *N* pixels. This, of course, assumes that the range values in that neighborhood are constant. Unfortunately, this is not true in most cases. The effect of averaging is more drastic if the averaging is applied on a spatial edge, or temporal edge in the form of motion. A filtering scheme that takes the edges into account gives more reliable results, especially around the edges. To accomplish this, one option is to use median filtering, where each pixel is assigned the median of the values around its local neighborhood. Another option is to apply edge detection prior to averaging. We evaluate the effects of various averaging schemes, such as median filtering and uniform averaging in Section 5.

4.2. Aliasing Problem

Due to the limited unambiguous distance range, aliasing arises as an issue. Target objects that are separated by one full range or integer multiples of the full range, are indistinguishable. For example, given a 50 MHz modulation frequency, whose unambiguous range is 3m, an object at 1m and another at 4m from the sensor produce the same range value. Lowering the modulation frequency to, say, 10 MHz increases the unambiguous range to 15m. However, the resolution goes down by a factor of 5, as seen from the resolution equation of Section 3.

One method of increasing the unambiguous distance range is to take two or more measurements, each with a different modulation frequency. Suppose two distance measurements are made with two frequencies f_1 and f_2. Let R_1 and R_2 be the maximum unambiguous ranges for these frequencies. With two measurements, the effective unambiguous range is increased to the LCM (least common multiple) of R_1 and R_2. This method is illustrated in Figure 7. Suppose that there is an object at 13m from the sensor. If we only use one modulation frequency of 25 Mhz, we can infer that the object is either at 1m or 7m or 13m or 19m. If we only use 18.75 Mhz for modulation, we can infer that the object is either at 5m or 13m or 21m. Combining the two results, we conclude that the object must be at 13m. Of course, this object would still be confused with another at 37m (13m+24m), but, compared to the one-frequency case, the unambiguous range is extended significantly.

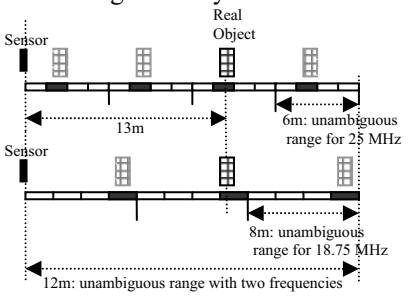


Figure 7. Ambiguity reduction with two different frequencies.

One strategy in selecting modulation frequencies is to maximize the LCM of the corresponding ranges by choosing two frequencies close to each other. Another strategy is to use a log-based approach where a number of measurements are made with frequencies *f*, 2**f*, 4**f*, 8**f* and so on. In each successive measurement the resolution is doubled while the range is halved. By

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $20.00 © 2004 IEEE
0-7695-2158-4/04 $20.00 (C) 2004 IEEE

combining the results, an accurate value of distance over a long unambiguous range can be obtained.

4.3. Motion Artifacts

As mentioned in Section 2, for higher accuracy, it is necessary to have the light signal switch between multiple phases within a single frame. However this approach may introduce motion artifacts. The effect is somewhat equivalent to having motion blur in a picture when a long exposure time is used. There is a similar motion artifact in our system if there is fast movement in the scene. When the object moves fast between the phases within a frame, each phase may potentially receive light coming from a different distance.

The motion artifact is observed mostly around the edges of a moving object. There are various ways to solve this problem. An obvious method is to increase the frame rate. Alternatively one can estimate the motion in 3D, and then correct the range values around the boundaries. Yet another option is to do edge detection followed by elimination of motion artifacts around the edges. Here, we propose a morphological method to solve the problem.

Let R_I be the input segmented foreground range image. Since the foreground is closer to the camera, the motion artifact is mostly observed in the foreground objects. This image is first binarized to obtain image B. The boundaries of the image can be extracted by an erosion operation. Let B_E be the eroded image. An eroded image R_E is obtained in the following manner:

$$R_E(r,c) = \begin{cases} 0 & if\ B_E(r,c) = 0 \\ R_I(r,c) & if\ B_E(r,c) > 0 \end{cases}$$

Then a dilated image R_D is obtained by a gray level dilation operation on R_E. Finally, the output range image R_O is constructed in the following manner:

$$R_O(r,c) = \begin{cases} 0 & if\ R_I(r,c) = 0 \\ R_D(r,c) & if\ R_I(r,c) > 0 \end{cases}$$

The output image R_O clears the motion artifacts. This method requires that the background be first eliminated via depth based segmentation. Then the background should be assigned a reference value of 0.

4.4. Ambient Effect

Ambient light is considered as unwanted light that has the same wavelength as the light source of the system. Although ambient light is non-modulated, constant light, it has a shot noise component that causes errors in the sensor. As a consequence, there may be noisy data in the area of the image receiving ambient light.

The ambient noise can be eliminated in various ways. In this paper, we propose a signal processing technique. We make use of the observation that the pixel values have a spatial (and temporal) high-frequency behavior in a local neighborhood when lit by ambient light. We first perform frequency transformation in a local neighborhood of every pixel. Let $F\{N[X_i]\}$ be the Fourier transformation where $N[X_i]$ is the set of pixel values around pixel i. We can then look at energy contained in the high frequency bands in $F\{N[X]\}$, and eliminate the pixel if this energy is sufficiently large.

In practice, this technique might be computationally too expensive. In order to make it efficient, one can use other filters such as:

$$Y(r,c) = \begin{cases} 0 & if\ f\left(N\left[X(r,c)\right]\right) > T \\ X(r,c) & if\ f\left(N\left[X(r,c)\right]\right) \le T \end{cases}$$

where $X(r,c)$ and $Y(r,c)$ are the input (original) and output (new) values at location (r,c) respectively, and $f(N[X(r,c)])$ is a function of the neighborhood of pixel $X(r,c)$. In our experiments, we use as $f(N[X(r,c)])$ the ratio of the standard deviation of pixels in the neighborhood of (r,c) to their mean value.

4.5. Saturation versus No Data Trade-off

This is a known problem for every camera based system. For instance, non-reflective objects may appear black while reflective objects saturate the pixels. In a TOF sensor, a pixel produces useful range data only when enough light is received at the pixel. On the other hand, the pixel produces a wrong range value when it saturates. Therefore, imaging low-reflectivity objects vs. high-reflectivity objects arises as a trade-off.

There are various techniques to increase the dynamic range. In this paper, we propose to use multiple exposure settings. In the low exposure setting, the high-reflectivity object returns enough light without saturating the pixel, and the low-reflectivity object does not produce any range value. In the high exposure setting, the high-reflectivity object saturates, while the low-reflectivity object produces the range data. Here we propose to combine the high-reflectivity object readings from the low exposure setting with the low-reflectivity object readings from the high exposure setting.

5. Experiments and Results

We have captured images of various scenes to test our ideas in resolving the issues mentioned in previous sections. First, we show some brightness and depth (range) images that were captured by our camera in Figure 8 and 9. In Figure 8, we image a large card board box that was tilted towards the camera. The brightness image does not provide enough information regarding the geometry of the box, but the range image shows the exact geometry. Here we used a colormap

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

such that the pixels change color from green to yellow and then to red as the objects are further away from the camera. We also provide a mesh view of the pixel values. In Figure 9, we show another scene where three blocks are placed on a platform. Besides the range image, the side view of a cloud map is also given to show the geometry of the blocks. In this example, we placed the blocks at 60cm, 74cm, and 94 cm respectively. The average reading on each of these blocks were 60.36cm, 74.11cm and 93.90cm respectively. The average standard deviation around a 5x5 neighborhood of each pixel was 0.88cm, 0.94cm, and 1.39cm respectively. We provide another example in Figure 10 where a person is imaged. We calculated the standard deviation around the local neighborhood of each pixel in the range image to be 0.54cm.

Next, we demonstrate the effect of median filtering in improving the resolution. A median filter of size 5x5 was applied to each pixel in the image of Figure 10. The resulting images are shown in Figure 11. The average standard deviation around all pixels has been calculated as 0.24cm. In other words, we observe a two-fold improvement in resolution using median filtering. We also applied uniform low-pass filtering across the image (Figure 11(c)). The average standard deviation was calculated as 0.22cm.

Another method for increasing the resolution is via temporal averaging. Figure 12(a) illustrates a face mesh captured by our sensor. The mesh is not smooth especially around the nose area. To improve the resolution, 10 consecutive images were averaged (Figure12.b). The details of the face are much clear and the resolution is considerably increased.

Next, we present experiments where we test the effect of ambient light and our filters for ambient artifact cancellation. In Figure 13, (a) and (b) show the brightness and range images of a scene where a small rectangular block is in front of a large rectangular block. Two incandescent lights of 600 watts each illuminate the scene. The small block is lit mostly by the light source of the system. The large block, however, is lit mostly by the incandescent light. In Figures 13(d) and (e) show, respectively, the small and large blocks in greater detail. In Figure 12(f) and (g), we provide the 2D frequency transformations of the two sub-windows. We observe that the frequency transformation of the large block has more high frequency components as an artifact of the ambient light. The ambient cancellation filter produces the image given in Figure 13(c). Note that the part of the scene lit mostly by the ambient content is eliminated.

Figure 14 shows a person sitting in front of the camera where the scene is lit by the same incandescent lights. Figure 14(c) provides the results of the ambient cancellation filter. In this example, the person is accurately distinguished from the background.

In Figure 15, we show one frame from a sequence where a rectangular block was captured in motion. We observe the motion artifact around the edges of the block in this example. In order to remove this artifact, we apply the morphological operations as described in Section 4.3. The result is given in Figure 15(c) where the motion artifacts are eliminated.

Finally, we illustrate the use of multiple exposures to increase the dynamic range. For this experiment, we created a scene where we placed high reflectivity objects close to the camera, and low-reflectivity objects away from the camera. We then obtained the brightness and range images with various shutter time settings. These images are shown in Figure 16. Observe that in each case, either the nearby objects saturate, or the more distant objects do not produce any range data due to lack of light. We combined the range measurements from two exposure settings, resulting in the images of Figure 17. We also obtained a brightness image by logarithmic scaling in between frames. We observed that a range image that shows the range values of all of the objects in the scene can be constructed with this method.

6. Discussion and Comparison to Other Types of Depth Sensors

Our depth sensor calculates distance by measuring the time that a light beam takes to travel from an object to the sensor. TOF techniques have been employed in radar systems for a long time. But our main contribution is to realize this technology using light (laser/LEDs), and implement it on a single CMOS chip. Due to the low cost of manufacturing CMOS chips, this technology is now available to a wide variety of applications.

In our experiments to evaluate the performance of our system, we observed a standard deviation of less than one cm. The resolution may be further improved by increasing the light power or by using post processing techniques such as median filtering. Theoretical analysis shows that depth resolution depends mostly on the amount of light reflected back into the sensor. There are various parameters affecting the amount of reflected light. For instance, if the light power, or the light exposure time is increased, the amount of the reflected light also increases, which in turn improves the resolution. Similarly, if the area illuminated by the light source is decreased, the amount of light received by each pixel is increased. The resolution can also be improved if the modulation frequency of the light source is increased.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04$20.00 (C) 2004 IEEE
0-7695-2158-4/04$20.00 (C) 2004 IEEE

Other sources of IR light in the environment, i.e., ambient light, may adversely affect the performance of the system. But there are various ways to improve the performance in these cases. In this paper, we proposed a method that applies texture analysis and removes the ambient content from the image.

TOF sensors have numerous advantages over other depth sensors. Triangulation-based methods such as stereo require intensive post processing to construct depth images. This is not necessary for the TOF sensor, and the post processing usually involves a simple table-lookup to map the sensor reading to real range data. Unlike triangulation systems, our sensor does not require a baseline between its optical components, or any sensitive alignment.

While the performance of many depth sensors depends heavily on the lighting conditions of the scene, our system has its own floodlight and does not require any light from the scene. Structured light systems usually require a good illumination contrast in order to recognize a certain pattern in the image. This translates to high power and high-density light sources, which are more expensive and potentially not eye safe. On the other hand our sensor doesn't require a high-power light source as there is no pattern to extract from the image. If enough light returns back to a pixel, the depth is successfully calculated.

Our TOF-based sensor has other advantages as well. It is texture independent, and constructs depth images regardless of the texture in the scene. Within a frequency-dependent range, it does not produce any ambiguous data. Finally, the sensor can be implemented on a standard CMOS chip, making it highly cost effective and commercially available in large quantities. Such a low-cost 3D sensor opens the door for many existing and new applications into the world of electronic perception technology where electronic devices are given the ability to perceive and interact with the world around them in real time.

7. Conclusions

The newly emerging electronic perception technology enables a large variety of present and future applications where a system or a robot needs to see and understand its environment. In this paper, we presented a new TOF-based camera system that generates depth images in real time for use in electronic perception applications. Our contribution is that we have integrated a complete TOF-based 3D depth sensor on a CMOS chip and developed the software methods to optimize its peformance. Although there are various problems involved in using such a system, we have shown that they can be successfully solved, resulting in a robust and efficient system with a depth resolution of only a few milimeters. We envision a wide variety of applications where our technology can be used to enable everyday devices to perceive and interact with their surroundings in three dimensions.

8. References

[1] S. T. Barnard, and W. B. Thompson, "Disparity analysis of images", *IEEE Trans. Pattern Anal. Mach. Intell*, 1980, 2(4), 333-340.

[2] K. L. Boyer, and A. C. Kak, "Color-encoded structured light for rapid active ranging," *IEEE Trans. Pattern Anal. Mach. Intell*, 1987, 9(1), 14-28.

[3] , J. K Aggarwal, and Y. F. Wang, "Inference of object surface structure from structured lighting -- an overview," *Machine Vision: Algorithms, Architectures, and Systems*, Academic Press, San Diego, CA, 1988, pp. 193-220.

[4] S. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell*., 1994, 16(8):824-831.

[5] A. Pentland, "A new sense for depth of field," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1987, 9:523-531.

[6] C.P. Jerian and R. Jain. "Structure from motion -- a critical analysis of methods," *IEEE Trans. on Systems, Man and Cybernetics*, 1991, 21(3):572--588.

[7] R. Miyagawa and T. Kanade, "CCD-based range-finding sensor," *IEEE Trans. on Electron Devices*, vol. 44, no. 10, pp. 1648 – 1652, Oct. 1997.

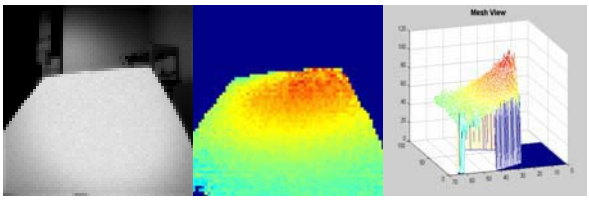
[8] 3DV Systems web pages: http://www.3dvsystems.com.

[9] C. Bamji, E. Charbon, "Systems for CMOS-compatible three-dimensional image sensing using quantum efficiency modulation," US Patent 6,580,496, granted in 2003.

[10] R. Jeremias, W. Brockherde, G. Doemens, B. Hosticka, L. Listl, P. Mengel, "A CMOS photosensor array for 3D imaging using pulsed lasers," *2001 IEEE International Solid-State Circuit Conference*, ISSCC 2001.

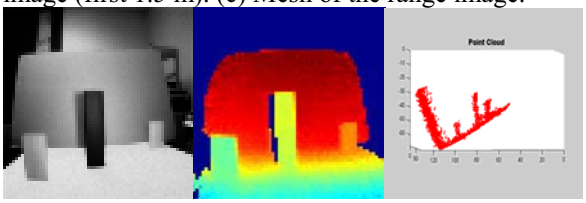
[11] M.D. Adams, "Coaxial range measurement – current trends for mobile robotic applications," *IEEE Sensors Journal*, Vol. 2(1), February 2002.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04$20.00 (C) 2004 IEEE
0-7695-2158-4/04$20.00 (C) 2004 IEEE



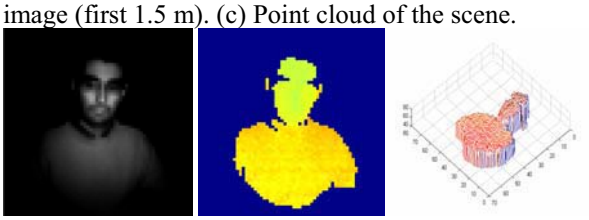
(a) (b) (c)

Figure 8. (a) Brightness image of a scene. (b) Range image (first 1.5 m). (c) Mesh of the range image.



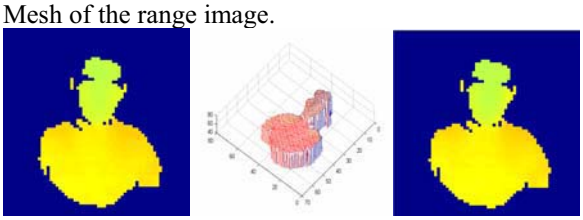
(a) (b) (c)

Figure 9. (a) Brightness image of a scene. (b) Range image (first 1.5 m). (c) Point cloud of the scene.



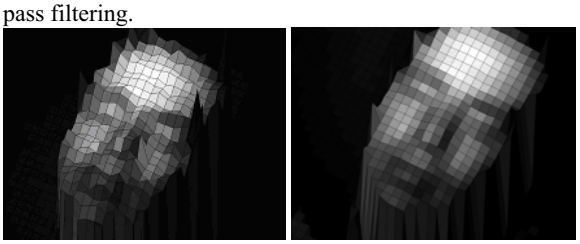
(a) (b) (c)

Figure 10. (a) Brightness image. (b) Range image. (c) Mesh of the range image.



(a) (b) (c)

Figure 11. (a) Range image after median filtering. (b) Resulting mesh. (c) Range image obtained with uniform low-pass filtering.



(a) (b)

Figure 12. (a) A face mesh obtained by our sensor. (b) The face mesh obtained by averaging 10 frames.

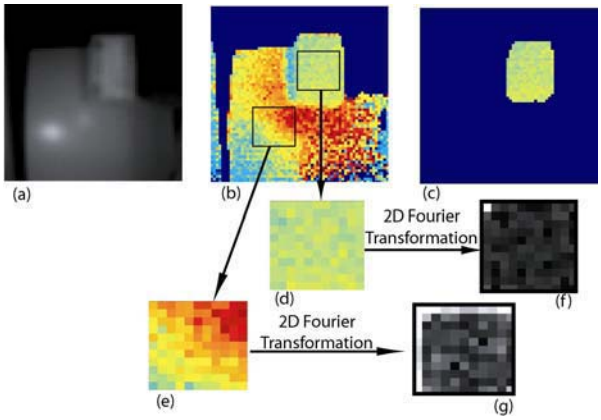
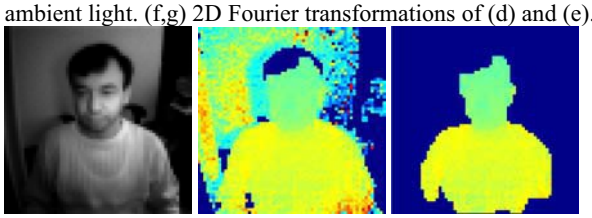
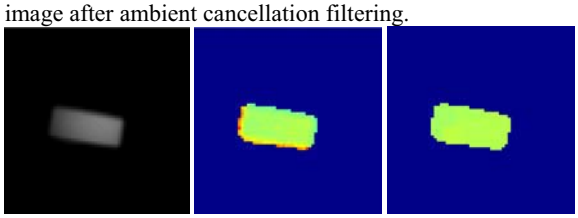


Figure 13. (a) Brightness image. (b) Range image. (c) Range image after ambient cancellation filtering. (d) Close-up of a region lit by the light source. (e) Close-up of a region lit by ambient light. (f,g) 2D Fourier transformations of (d) and (e).



(a) (b) (c)

Figure 14. (a) Brightness image. (b) Range image. (c) Range image after ambient cancellation filtering.



(a) (b) (c)

Figure 15. (a) Brightness image of an object in motion. (b) Range image. Observe the motion artifact along the edges. (c) Range image after morphological motion filtering.

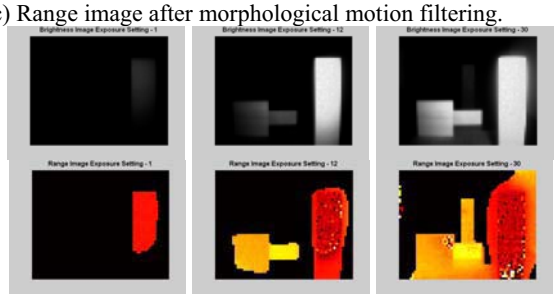


Figure 16. Brightness (top row) and range (bottom row) images using various exposure settings.

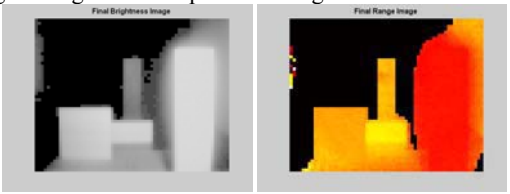


Figure 17. Brightness (left) and range image (right) obtained using two exposure settings.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04$20.00 (C) 2004 IEEE
0-7695-2158-4/04$20.00 (C) 2004 IEEE