A New NIR Camera for Gesture Control of Electronic Devices

Dan Ionescu, Viorel Suse, Cristian Gadea, Bogdan Solomon
School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Canada
{dan, viorel, cgadea, bsolomon}@ncct.uottawa.ca

Bogdan Ionescu, Shahidul Islam
Mgestyk Technologies, Inc.
Ottawa, Canada
{bogdan, shahid}@mgestyk.com

Abstract—Since the introduction Gesture Control technology in the electronic gaming technology a series of attempts have been made to deploy it also on other domains such as robotics, teaching, medical, automotive and many others. Human gesture used for Man-Machine Interaction became attractive as it offers a simpler way of controlling sophisticated devices, in a sci-fi-like scenario, in return of an increasingly computational power required by the artificial intelligence algorithms needed to detect, track and recognize them. There have been attempts to bring a solution to it by using 2D or 3D based image processing methods. There is a clear balance incline towards 3D methods in the consumer product as besides the almost insurmountable difficulties for producing robust and stable results, the price constraint added supplementary hurdles. As perfect illumination conditions are core factors in obtaining the above results, the infrared light was unanimously adopted by the domain technologies.

In this paper, a novel real-time depth-mapping principle and a corresponding hardware solution for an IR depth-mapping camera is introduced. The new IR camera architecture comprises an illuminator module which is pulsed and modulated via a monotonous function using a phase-locked loop control for the laser intensity, while the reflected infrared light is captured during the increasing and decreasing monotonous function. A reconfigurable hardware architecture (RHA) unit calculates the depth and controls the IR waves in synchronism with the infrared sensor. The resolution of the depth map is variable depending on the resolution and gating possibilities of the image sensor. A sensor of 1 megapixel is used, providing a resolution of 1024x1024. Images of real objects are reconstructed in 3D based on the data obtained by the laser controlled by the RHA. A corresponding image processing algorithm builds the 3D map of the object in real-time. In this paper the camera is used to control consumer electronic products such as TV sets, laptops and others.

Keywords—3D camera technology; real-time 3D object reconstruction; infrared cameras; gesture control; virtual environments

I. INTRODUCTION

Gesture Control was revived in late 2008, when a new way of interacting with computers was demonstrated [2] using a special 3D IR camera and OpenCV. A series of image processing algorithms were created to make computers understand gestures represented by hands and finger movements. The advent of Kinect technology, adopted by Microsoft from PrimeSense, consolidated the Gesture Control as a new paradigm for the man-computer interaction. Applications have mushroomed and eventually a series of works [3] have been dedicated to a deeper analysis of the technology.

A camera which operates using an image sensor such that its output contains depth information has been the key element in obtaining the well-known six degrees of freedom in user interaction [8]. A variety of computer applications and electronic devices such as laptops, TVs, set-top boxes, and others are in search for gesture based control and are open fields for using depth cameras as the only element capable of providing a natural and intuitive interface. Applications range from 3D TV to computer gaming, animation, biomedical imaging, and robotic vision, just to name a few.

3D Imaging has been a domain on its own long time before gesture control has been revived. The field was known as the 3D object reconstruction and methods to realize it can be categorized into 3 basic techniques: i) stereo or triangulation-based, known as stereo-vision which implies measuring geometrical attributes in order to extract the range data; ii) structured light techniques, where depth images are obtained by projecting a known pattern of light and recording the deformations of the pattern relative to a reference image which contains the pattern projected on a planar surface; iii) LIDAR and time-of-flight-based imagers, which obtain a depth map through scanning mechanisms that measure the time or the phase taken for a generated pulse of light to return to the camera.

Stereo vision principles have been applied using multiple pairs of cameras, where overlapping sections of images can be used to obtain three dimensional data as long as there is enough detail present for matching [4]. Depth reconstruction from 2D images was discussed in a vast amount of literature and has captured the attention of researchers since late eighties [4]-[7]. Each method varies over the others in terms of precision, costs, computational requirements and advantages or in terms of what it can and cannot be reconstructed. Barnard and Fischler [4] along with Dhond and Aggarwal [5] analyzed some of the methods used for stereo fusion and produced valuable surveys.

Structured light is a well-established technique which is used to obtain three dimensional (3D) information of a scene.
from recorded 2D images, one of which can be the structured light projector. These methods use a specially designed light source to project sheets or beams of light with a known a-priori spatial distribution onto the scene casting lines [10] or points (dots) [9] on the objects of interest. The major advantage of this technique is that it circumvents the natural discontinuity/monotony properties of the object surfaces by replacing them with artificial features of structured light projections. These are better recovered and interpreted by computer vision. Salvi [11] wrote a report on different codification strategies available in structured light systems. More recently, the Kinect camera used near-infrared light (NIR) to project a pattern of dots and observe their shift relative to the reference pattern, producing depth images at a maximum resolution of 640x480 [3].

The time-sensitive gating techniques required by LIDAR and TOF systems are typically implemented using image-intensifiers [6], [12]-[15] or other exotic components such as photovoltaic cells, or very fast (ns) shutters. In this paper, a new system for real-time measurements of depth based on variable gain and gating techniques which is called the space slicing effect is introduced. The novel approach devised opens new directions that eliminate the need for using costly and/or exotic devices such as image intensifiers, while also supporting high image resolutions available from off-the-shelf image sensors, only one of which is required. The space slicing principle is based on modulated near infrared light whose intensity is controlled through the process of building basic images which are processed based on an image processing algorithm for calculating depth-information. The method also uses a relatively high frame rate CMOS sensor which performs the gating so that the desired depth precision is obtained.

II. THE 3D IMAGING METHOD

In this new method, the basis of the depth measurement is accomplished by generating light pulses, by a module called the illuminator of the camera, with high frequency, a variable duty cycle, and variable frequency. Thus a cone of near-infrared (NIR) is projected onto the scene. For the camera described in this paper, the infrared light was chosen to be in the 850-900nm range, while the total pulsed power was set to 150 mW to obtain the desired maximum depth distance of 5m (the light is dispersed upon leaving the illuminator). Living room-based gesture interactions typically require people to be located at a variable distance from one to f meters such that either fingers or the whole body can be acquired properly by the combination of optics and the image sensor. The light source is generated as a square laser pulse of short and variable duration. The infrared light produced is an expanding spherical surface of finite width determined by the square pulse of the laser module (a “light wall”). The wall is also controlled in intensity and a special intensity function is generated based on a monotonic pulse width modulation (PWM) function implemented by the camera controller program. The infrared light is reflected back to the camera by the real-world 3D scene, and provides information “slice by slice”, which is translated into a depth image by a depth image video processor.

Figure 1 shows the high-level design of the camera. Two boxes are observed as objects and the corresponding depth image is obtained on the PC via USB. The electronics of the camera contain the following main blocks: i) an infrared light generator, called the illuminator, ii) an image sensor, and iii) a camera controller board containing the video processor.

The duration of the illumination, which is proportional to the distance of the object to the camera, is controlled by the video processor. The illuminator therefore sends controlled pulses to the object, thereby exploring it in a spatial manner. By repeating this slicing process a number of times and by controlling the camera parameters in real-time, a depth image of the object will be produced. The real-time nature of the camera requires that a new depth image be obtained every 33.3ms (30fps), which minimizes the delay between the real-world object movement and the frame received by the media processor which processes the images frame by frame in real-time.

One of the most important features of the depth camera is the ability to change the parameters of the depth measurements on an as-needed basis. It can therefore be adjusted to process nearby hand gestures, distant full-body movements, or both, as needed by the application. This is an important feature as presently there is no other such camera which can be used for finger detection. The camera can therefore produce depth images which will include certain objects, while disregarding others before or beyond the limits of the defined “wall”. This is made possible by the novel “space slicing” principle, which is used to build the depth-map of the objects present in the field of view (FOV).

Figure 2. Space slicing principle.

The “space slicing” is based on the following space exploration technique: a cone of an IR light wave is
generated by a controlled and dispersed laser illumination source. The illumination is modulated in duration and frequency according to an adaptive algorithm such that it explores slices of the object in the FOV in sync with the frame rate of the sensor. The modulation algorithm is calculated by a video processor implemented using a reconfigurable hardware architecture unit, which also implements the depth calculation from which the feedback for the adaptive control is taken. This IR cone wave modulation is illustrated in Figure 2, where the pulses from the illuminator in Figure 2A reflect off of the boxes whose images are eventually generated by the camera as in Figure 2B.

Using elementary image processing operations, such as a special thresholding technique and image subtraction and reconstruction the pixels contained by the space slices are merged eventually in a depth image providing (x,y,z) coordinates for each of them. The algorithm will produce using slice by slice, two images (a primary and a secondary) which are further used to calculate the depth by using the formula in (1).

\[
d = \frac{vT_s I_+}{4(I_+ + I_-)} - \frac{v t_p}{2}
\]  

Equation (1) contains the following symbols:
- \(d\) = Depth of the pixel
- \(v\) = Speed of light in free space = \(3 \times 10^8\) m/s
- \(I_+\) = Intensity in the primary frame
- \(I_-\) = Intensity in the secondary frame
- \(T_s\) = Intensity modulation period; for example \(T_s = 250\) ns = \(250 \times 10^{-9}\) s (adjustable)
- \(t_p\) = Time offset; for example \(t_p = 60\) ns = \(60 \times 10^{-9}\) s (adjustable)

Figure 4. Resulting depth slices.

For example, with the formula in (1), the images of the three cubes on the right side of Figure 2 become the shaded areas in Figure 4.

Figure 5. Depth data makes fingers distinguishable.

In reality, the depth map can be calculated to obtain 8 or 16 bit pixel values, which can be represented either as grey levels or using artificial colors.

Figure 6 summarizes the electrical design of the camera’s “depth engine” and how the various components interact to produce a depth image. The Depth Engine is composed of the Depth Engine Video Processor, a DDR2/DDR3 Memory Controller, a USB Controller, a SPI Controller, a Command Decode and Execute Module, an Illumination Controller, and a DSP Microcontroller Interface. In the figure, “D” is used to signify the exchange of “data”, while “C” represents “control” and “S” represents “status”.

The Depth Image Video Processor receives the space sliced images from the image sensor in groups of eight or four depending on the camera tuning used. All the depth calculations take place in parallel in the hardware engine. The hardware engine controls the illuminator and keeps the synchronization between the image sensor and the illuminator. The control takes into account the synchronization necessary for the production of the primary and secondary images such that the depth image can be calculated by \((1')\). After the distance from the camera to the object is calculated the resulting depth images are mapped into a grey (8 bits) or pseudo colored image (16 bits). The resulting precision of the depth calculations will correspond to the selection of one or the other of the mapping. In the same engine resides also the algorithm for face detection which is the starting point of the skeleton detection algorithm.

The resulting image is then processed by the finger/hand or full-body detection, recognition and tracking software algorithms. The algorithms are implemented by the Digital Media Processor such that the 3D camera exports only the controls which correspond to the specific gestures to which the application is responding.

III. EXPERIMENTS AND RESULTS

The camera was implemented as described and a variety of experiments were performed on the resulting data. Figure 3 shows the various steps taken by the camera while building
the depth-map of a box being held at an angle. Figure 3a and Figure 3b show the primary and secondary images, respectively, as obtained by the camera CMOS sensor at two different settings. Figure 3c shows the final depth image visualized as a colored image, while Figure 3d shows the obtained depth as a grayscale image. The corners of the box that are further from the camera have a different grayscale value than those closer to the camera.

In this situation, the distance of 1m was measured with an error of 1.5 mm. This is sufficient for applications in which hand and leg movement has to be tracked [16]. This amount of error is for robustly tracking finger movements.

The new 3D camera brings the advantage of a considerably high resolution of 1024x1024 based on the image sensor that was used, although the principle will also function with much higher resolution sensors. This helps to provide detailed and stable data, which is important for gesture processing. Additional benefits to gesture processing include the ability to focus all processing on the specific “slice” that contains the relevant data (such as the fingers), as well as the ability for the camera to operate independently of lighting conditions (such for controlling a TV in a dark living room).

Figure 5 and Figure 7 show the result after camera illumination timings were tuned to find a middle-ground in which both finger data and full-body data were visible. The fingers on the hand are clearly distinguishable and are on a different depth layer than the user’s body. Background objects are eliminated.

The new 3D camera brings the advantage of a considerably high resolution of 1024x1024 pixels based on the image sensor that was used, although the principle will also function with much higher resolution sensors. This helps to provide detailed and stable data, which is important for gesture processing. This also allows the camera to operate independently of lighting conditions (such for controlling a TV in a dark living room). Additional benefits to gesture processing include the ability to focus all processing on the specific “slice” that contains the relevant data (such as the fingers).

Two objects, one white with high reflectivity and one gray with medium reflectivity, were positioned at different distances from the camera and the resulting depth pixel value was measured for each distance. The results were plotted in Figure 8. From this data, it was calculated that the camera offers a depth resolution of approximately 1.5 cm at a distance of 3m.

![Figure 6: Block diagram of 3D camera depth engine.](image-url)

![Figure 7: Full-body and hand visible at different depths.](image-url)
IV. APPLICATIONS

The 3D camera can be used in any computer application whose control through gestures or full-body motion makes sense. For example, by using image processing algorithms to detect and track finger movements, a user is able to use gestures similar to multi-touch surfaces to pinch and grab a 3D object within a virtual environment. Such an application was implemented to evaluate the accuracy of the gesture-based control, as shown in Figure 9.

In this application a gesture based finite state machine is used to allow a user to manipulate an anatomic object. The 3D model of a human brain can be rotated around the x, y, and z axes, can be translated to any point within the FOV of the camera, and can be decomposed into its anatomic components. The example illustrates the use of gesture control in a virtual environment dedicated to teaching.

The electronic gaming area is yet another domain in which gesture-based control can be used. Figure 10 shows the user controlling an existing game on the Sony PlayStation III video game console. The gesture control camera presented in this paper can be used with any game as the camera sends the controller commands which the game console is responsive to.

Similarly the gesture control implemented by the camera can be used to control a TV or set-top box with gestures instead of using the ever-disputed remote.

V. CONCLUSION

The recent trend towards motion and gesture-based man-machine interfaces has created a demand for 3D cameras which can make such interfaces a reality. Existing 3D cameras require expensive components or can only produce depth images of relatively low resolutions. This paper has introduced a new infrared-based 3D camera and its reliable method for acquiring 3D depth data based on a novel “space slicing” technique. The real-time acquisition of the high-resolution depth image, as well as its interpretation and understanding, signal the turning point of a new era in human-computer interfaces, with applications in numerous fields, including education and medicine. This was illustrated with a 3D brain application where test users successfully manipulated the virtual object by using finger gestures. Future research will test the camera for the control of electronic devices such as TV sets, set-top boxes, laptops, and many others. In addition, a more thorough comparison of the depth characteristics relative to other 3D camera technologies will be performed as the camera continues to be improved.

REFERENCES


