16

Structured Light Illumination Methods for continuous motion hand and face-computer interaction

Charles J. Casey, Laurence G. Hassebrook and Daniel L. Lau
University of Kentucky
Department of Electrical and Computer Engineering
Lexington, Kentucky
USA

1. Introduction

Traditionally, human-computer interaction (HCI) has been facilitated by the use of physical input devices. However, as the use of computers becomes more widespread and applications become increasingly diverse, the need for new methods of control becomes more pressing. Advances in computational power and image capture technology have allowed the development of video-based interaction. Existing systems have proven themselves useful for situations in which physical manipulation of a computer input device is impossible or impractical, and can restore a level of computer accessibility to the disabled (Betke et al., 2002). The next logical step is to further develop the abilities of video-based interaction. In this chapter, we consider the introduction of third dimensional data into the video-based control paradigm. The inclusion of depth information can allow enhanced feature detection ability and greatly increase the range of options for interactivity. Three-dimensional control information can be collected in various ways such as stereo-vision and time-of-flight ranging. Our group specializes in Structured Light Illumination of objects in motion and believe that its advantages are simplicity, reduced cost, and accuracy. So we consider only the method of data acquisition via structured light illumination. Implementation of such a system requires only a single camera and illumination source in conjunction with a single processing computer, and can easily be constructed from readily available commodity parts. In following sections, we will explain the concept of 3D HCI using structured light, show examples of facial expression capture and demonstrate an example of a “3D virtual computer mouse” using only a human hand.

2. Structured Light Illumination

Structured light illumination (SLI) allows one to measure the depth information of a surface by measuring the deformation in a projected light pattern (Schmaltz, 1932). A simple example would be a pattern of stripes projected onto a sphere. When viewed obliquely, the light stripes on the sphere appear curved as shown in Fig. 1. For a given arrangement of the projector and camera, the variation in a pattern can be characterized extremely accurately, such that a precise model of the surface can be reconstructed. Most modern implementations of SLI systems make use of digital projectors to illuminate the subject and a digital camera to capture an image of the illuminated subject, though in certain cases static projection devices (slide projectors, for example) may be used.

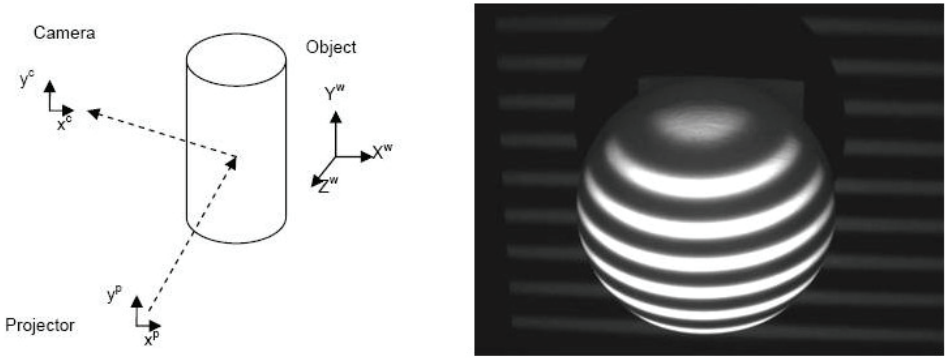

Fig. 1. (left) SLI geometry and (right) example stripe pattern on sphere.

Mathematically, the SLI measurement process is based on triangulation. Accurate results can be produced only when there is a well defined relationship between a single point on the projection plane and the corresponding point on the captured image, as shown in Fig. 1. It is to establish this relationship that projection patterns are utilized. A projection pattern (or more frequently a series of patterns) is designed such that each pixel (or row or column, depending on the specific implementation) of the projection image is uniquely characterized, either by some characteristic intensity value sequence or by some other identifiable property such as a local pattern of shapes or colors. When projected onto a subject, the captured image (or series of images) can be analyzed to locate these identifiable projection pattern points. Given a fixed placement of camera and projector, the location of any given pattern point on the subject creates a unique triangle with dimensions defined by the depth of the subject surface.

A well designed projection pattern can achieve significant accuracy and precision (Li et al., 2003). However, due to the difficulties involved in encoding each pixel uniquely, many of the most effective pattern types require more than one projection/capture instance in order to reconstruct the subject surface. These are known as time-multiplexed patterns. For example, the Phase Modulation Profilometry (PMP) (also known as sine wave shifting) technique utilizes a series of patterns in the form of sinusoidal grayscale gradient images, each shifted by a certain phase angle. Varying the intensity in this way will encode each row (or column) in the projector with a unique phase value. However, it has been shown that the accuracy of the phase value (and thus, the measurement of subject depth) is dependent on the number of shifts used. Therefore, a more accurate surface reconstruction requires more pattern projections and thus a longer scanning process.

2.1 Structured Light Illumination Motion Capture Methods

For many scanning applications it would be ideal to capture the required surface information in only a single projection/image instance. This is because during a multiple pattern scan, subject motion introduces error into the depth measurement. These techniques are therefore largely problematic for motion capture and human interaction. Fortunately, there are other options for single frame SLI capture, which offer not only reduced scan time, but also allow one to scan a moving subject. Most single pattern techniques fall into two categories; color-multiplexed types and neighborhood-search types.

Color-multiplexing simply combines individual patterns from some multi-pattern technique into a single pattern by coloring each differently. A three pattern PMP sequence, for example, can easily be combined into a single pattern by coloring each of the three patterns red, green, or blue. In this way, each pattern can be isolated independently of the others by considering only the R, G, or B channel of the captured image. Each channel image is effectively identical to a single frame of the corresponding multi-pattern PMP scan process. While the concept is simple, the number of patterns that can be combined in this way is usually relatively limited and analysis is plagued by non-idealities (Pan et al., 2006). In addition, color patterns introduce a strong dependence on subject coloration and luminance properties. If a subject is strongly colored blue, for example, there may be insufficient information in the R and G image channels to properly reconstruct the surface.

Neighborhood-search methods take a different approach entirely. These techniques utilize a pattern (usually binary in nature, that is, black and white colors only) in which subsections of the pattern can be uniquely identified in some way. Specific implementations may utilize patterns of noise or streaks (Maruyama & Abe, 1993) in which a point can be identified according to the known local statistical characteristics of the pattern, or deterministic sub-patterns defined by M-arrays or De Bruijn binary sequences (Morita et al., 1988) (Hall-Holt & Rusinkiewicz, 2001) wherein the identity of one point can be determined by the information contained in nearby points. Like color-multiplexed systems, the accuracy of neighborhood-search based techniques may be strongly dependent on subject surface characteristics. If pieces of the pattern can be obscured by subject features or distorted too much by local gradients, correct identification of the pattern points may be impossible. In addition, the primarily binary nature of the patterns can limit the resolution possible. Thus, only a small portion of surface points may be measured.

There are other methods of motion capture based on entirely different concepts. One method is to utilize high-speed hardware to simply run multi-pattern sequences faster (Zhang & Huang, 2006) reducing the effect that the subject's movement has on the depth measurement to a negligible amount. The composite pattern technique (Guan et al., 2003) combines component patterns of a time multiplexed method, such as PMP, into a single pattern by modulating each by a known frequency. This allows an effect similar to that of color-multiplexing, but avoids many of its drawbacks. Another new method, Lock and Hold (Hassebrook & Lau, 2006), utilizes a multi-pattern scan followed by a continuously projected tracking pattern (the Hold pattern) in order to scan moving subjects. The latter two options advantageously require no specialized hardware.

2.2 Composite Pattern

SLI systems frequently utilize patterns that vary only in a single direction (the "phase direction", a term taken from PMP and "y^p" in Fig. 1) and are constant in the other (the "orthogonal direction" or "x^p" in Fig. 1). In such a system, the camera is offset from the projector along the phase axis only. In this way, depth variation will cause variation in the pattern along the phase direction while leaving the pattern unaltered in the orthogonal direction. To visualize the effect, consider a square projected onto the surface of a sphere. If one views the sphere from an offset parallel to one set of sides of the square, these sides will still appear straight, while the other sides will appear curved. Composite Pattern multiplexing takes advantage of this fact by introducing sinusoidal variation along the orthogonal direction. When a surface illuminated with a composite pattern is viewed by the camera, the modulating signal will be unaltered by the surface features and can be used to isolate the component patterns in a way analogous to isolating each channel in an RGB color-multiplexed image.

Consider a four pattern PMP sequence. The intensity of each of the four patterns varies sinusoidally along the phase direction only. To combine the four into a single pattern, each is element-wise multiplied by a modulating image; patterns which vary sinusoidally in the orthogonal direction only, each at a unique (relatively high) frequency, as shown in Fig. 2. The modulated patterns are combined (and the resulting intensity scaled as necessary for the projection device) to create the composite pattern.

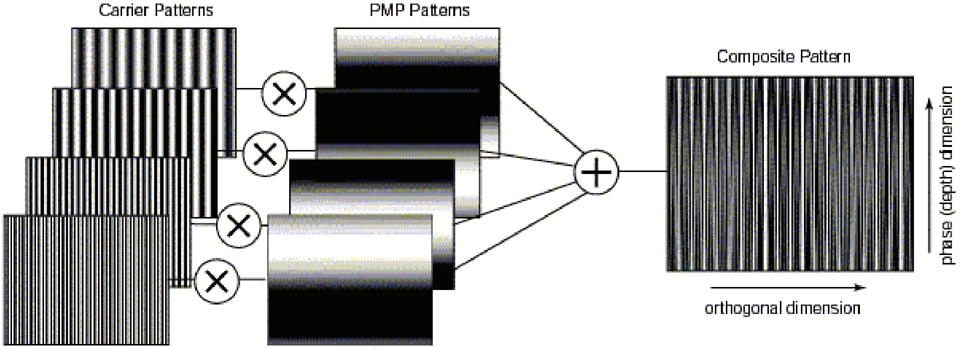

Fig. 2. Composite Pattern combining 4 PMP patterns into a single composite pattern.

When a composite pattern is used to illuminate a subject, the captured image must be processed in order to extract the original component PMP pattern images. These images can then be analyzed as though they were individual frames in a multi-pattern scan sequence. The process of isolating the component images is very similar to the process of isolating modulated communications channels. Considering the 2D Fourier transform of the image, component pattern information will appear as four signal envelopes shifted in the orthogonal direction by the modulating frequencies. Each pair of these envelopes (considering both positive and negative frequencies) is isolated using 2D band-pass filters. The inverse Fourier transform of these isolated bands are the equivalent component pattern images, and are then used to determine the surface depth according to standard PMP methodology. The method was combined with correlation filters to track hands (Guan et al., 2003) and used to control a virtual reality point of view as shown in Fig. 3. The left and right composite image of Fig. 3 have three component images; (upper left) the captured image, (upper right) the 3-D segmentation and hand tracking and (lower) the point of view of a virtual reality.

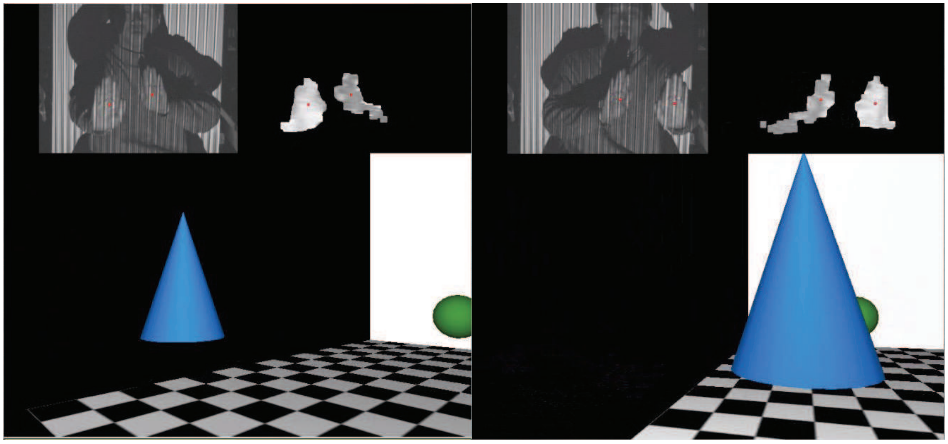

Figure. 3. (left) Rotated point of view by having one hand in front of another. (right) Translation of virtual reality to the left of the operator.

2.3 Lock and Hold Structure Light Illumination

Like the Composite Pattern technique, Lock and Hold motion capture was an idea inspired by communications theory. The idea is that, as in the operation of a phase-lock loop, if one can "lock on" to an unknown signal, then the changes in that signal can be easily tracked by compensating for the small incremental changes that occur through time. Lock and Hold motion capture uses an un-coded structured light pattern (usually a pattern of stripes with triangular cross sections) to capture the depth data of a moving surface. Changes in this "Hold pattern" are traced through the multiple frames of the capture video sequence in order to acquire a continually updated accurate depth map of the subject. Unlike similar systems that utilize un-coded SLI (Rodriguez et al., 2007) the system avoids difficulties involved with pattern ambiguity by the use of the "Lock sequence"; a preliminary 3D scan taken before the subject is allowed to move. Since an un-coded pattern has numerous identical elements, it can't generally be used to measure absolute depth in the same way as a coded pattern method (such as PMP or even Composite Pattern) since a projected pattern point may correspond to any number of pattern points on the captured image. By performing a preliminary 3D scan using PMP, the relationship between an identified point on the Hold pattern projection and Hold pattern capture can be unambiguously defined.

A simple explanation of the Lock and Hold process is as follows: to begin, a standard 3D scan of a subject is taken using a method such as PMP. Immediately following this, the Hold pattern projection begins and the subject is allowed to move, as shown in Fig. 4 (left and right). The Lock scan creates an unambiguous "phase map" which relates each point of the projection pattern to a single point that it illuminates on the subject image. If a Hold pattern is immediately projected, the first frame of the Hold capture sequence is directly related to the phase map. In other words, each isolated feature of the Hold pattern maps to a single phase value from the PMP scan. In this way, the depth of each isolated Hold pattern feature (i.e., “snake”) can be calculated using triangulation techniques.

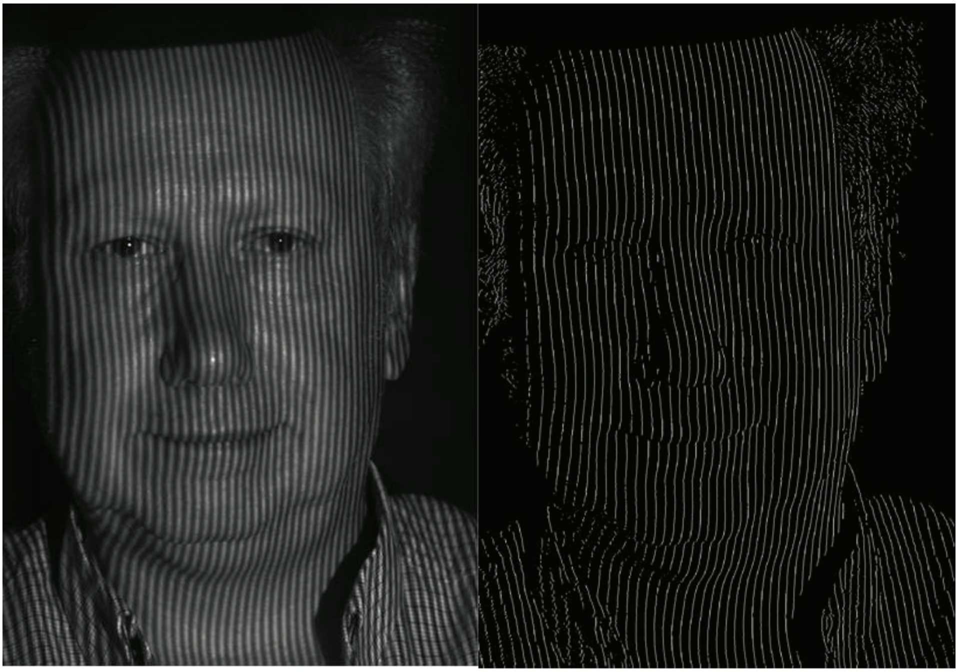

Fig. 4. (left) Example of a Hold pattern projection and (right) resulting “snakes” representing depth of Hold image.

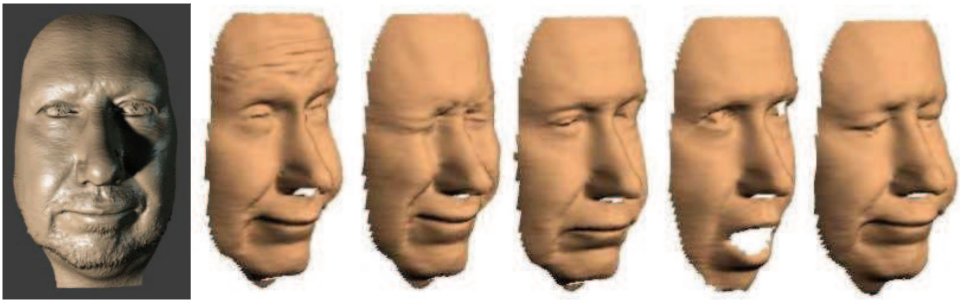

Fig. 5. (left) Lock scan and (right) sequence of Hold scans.

Once each feature in the first Hold frame is unambiguously identified, features in the next frame are isolated. Then, at each identified point in the first frame, a search is performed in a window around the corresponding position in the next frame. If a suitable feature is found in that frame, it is assigned the appropriate identity. In this way features can be traced through the numerous frames of the Hold sequence, and a depth map for each frame can be calculated. A Lock scan and 5 samples of Hold scans are shown in Fig. 5.

Practical implementations of the process require additional steps, of course. Depending on the shape of the subject and the speed of its movements (relative to the capture rate of the camera), the initial tracking process may identify some features incorrectly. Thus, techniques for error prevention or correction are normally required for optimal results. However, for our purposes, a detailed description of this process is not necessary.

3. Augmented Reality 3D Computer Mouse Demonstration

In subsection 2.3 we demonstrate that surface details can be obtained by SLI. The Lock and Hold method was designed for special effects applications where a high resolution Lock scan is needed as well as a series of lower resolution Hold scans. The Lock scan takes about 1 to 3 seconds to capture and could be replaced by a method we call “leading edge lock” where the object enters the Field of View (FOV) and the leading edge is used to acquire a non-ambiguous measure of depth and thus, lock the snakes to an initial depth for tracking during the Hold process. However, for the convenience of demonstration we use our existing scanner to show feasibility for using a hand as a interface device to the computer. To do this, we must be able to track a hand feature such as a finger tip. We will use a simple correlation filter to conduct a five finger tip tracking operation and use the position and depth of the fingertips to convey control parameters to the computer. The value of this control is limited by the accuracy of the fingertip position measurement so we provide a final experiment to obtain the accuracy of the depth position of a finger.

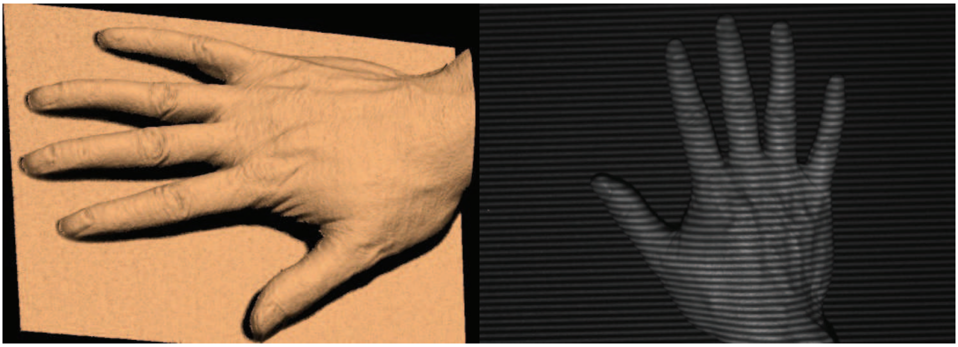

Fig. 6. (left) Lock scan of hand and (right) sample Hold image.

3.1 Fingertip Detection

The fingertip detection is accomplished globally by using a correlation filter designed to detect fingertips and suppress other regions of the hand. The Lock scan and a sample Hold scan are shown in Fig. 6 left and right, respectively. The captured image of the hand is down sampled to a course image for numerical efficiency has shown in Fig. 7 (left). That image is then correlated with a fingertip correlation filter leaving the detected fingertips as shown in Fig. 7 (middle). From those locations, the fingertip geometry is analyzed as a constellation of points to verify that they are actually fingertips.

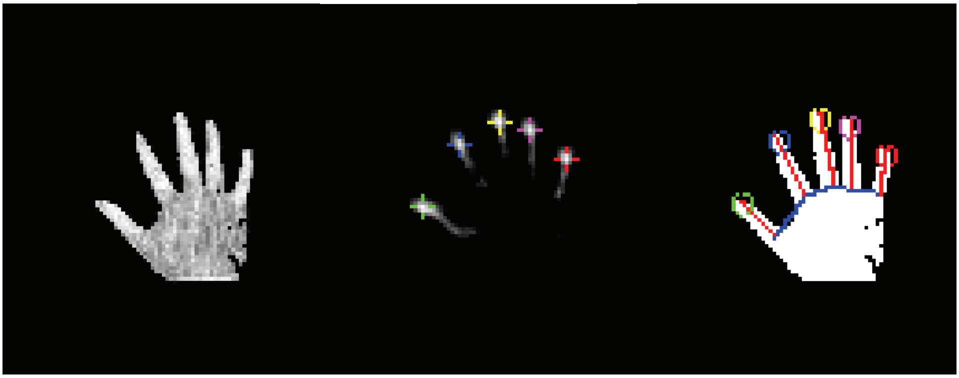

Fig. 7. (left) Course image of hand, (middle) correlation response and (right) characterized fingertip constellation.

The fingertip correlation filter is designed to both detect the fingertips as a circular region and also suppress non-circular shapes. As shown in Fig. 8, the filter has a circular region of positive values surrounded by a ring of negative values.

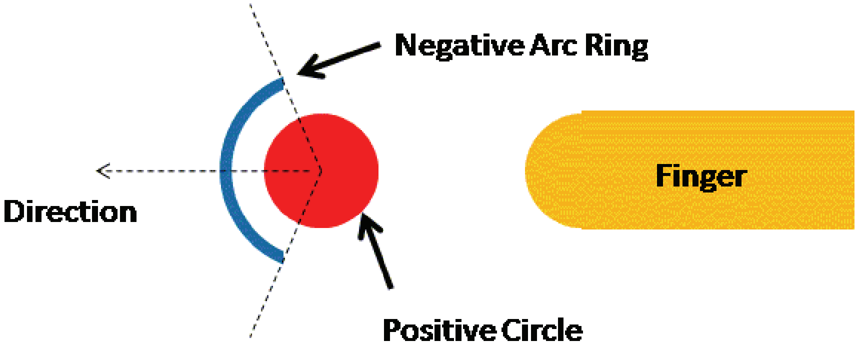

Fig. 8. The correlation fingertip filter.

The negative ring does not extend completely around the positive circle else it would partially suppress the tips. The positive and negative values are chosen such that when correlated with constant image intensity, such as the palm area, the resulting correlation is zero as shown in Fig. 7 (middle). Note that the filter only partially suppresses the fingers leaving the tips as the maximum correlation points. The number and positions of the points are checked to ensure their constellation is representative of a human hand configuration. Once the tip locations are established, the associated world coordinates, $\{X_w,Y_w,Z_w\}$, are used as the interface controls.

4. Results

We present two experiments and a discussion of numerical efficiency. The first experiment shows the tracking results for all 5 fingertips on a hand. The second experiment measures the depth accuracy of the fingertip position. The experiments were performed using an existing scanner system our group developed for a special effects application and then processing the data off line with a fingertip tracking algorithm used for biometric applications. So to evaluate the potential for a practical non-contact interface, we provide a discussion of numerical efficiency.

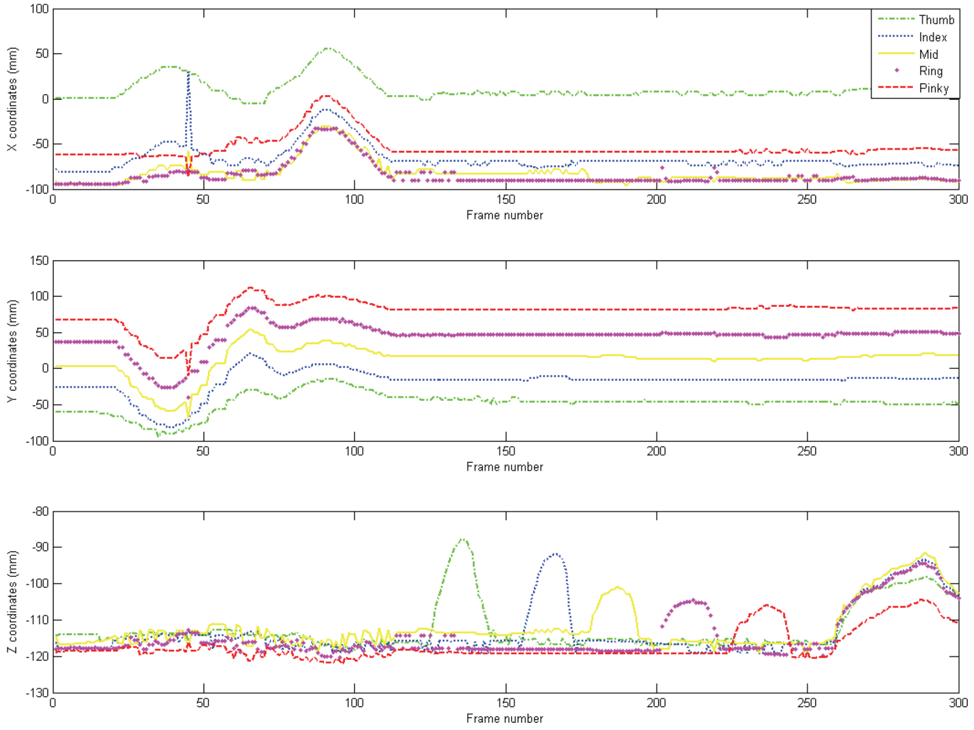

Fig. 9. XYZ tracking results of hand and finger movement.

4.1 Experiment 1: Five Fingertip Tracking

The fingertips were tracked for 300 frames at 30 frames per second (fps). The fingertip positions are shown in Fig. 9. The hand was moved in succession; left, right, back, and then forward to its starting position. Then, each finger, starting with the thumb, was raised up in sequence and finally all fingers were raised at the same time. The positive X axis in Fig. 9 is downward toward the wrist shown in Fig. 7. The positive Y axis is orientated toward the "pinky" finger side shown in Fig. 7. The positive Z axis direction is up off of the plane. A discontinuity occurred just before frame 50 and happened when three fingertips lost lock for one frame as shown in Fig. 9. The finger data is encoded by color such that the thumb (d1) is green, the “index” finger (d2) is blue, the “middle” finger (d3) is yellow, the “ring” finger (d4) is violet, and the “pinky” finger (d5) is red. The displacement of the thumb along the X axis, back towards the wrist, can be seen in Fig. 9 (top). The first hand movement is to the left as indicated by the Fig. 9 (middle) Y axis at about frame ~40. Note there is a rotation of the hand which also affects the X axis movement. The hand is then moved right to its maximum position at frame ~65. The hand is then moved back toward the wrist direction in positive X direction at frame ~90 and then returned to the original position at frame ~115. Next, starting at frame ~125, each finger is raised in the Z direction starting with the thumb (d1) followed by d2, d3, d4 and d5 ending at frame ~245. The last movement is the raising of all 5 fingers between frame ~260 through 300.

4.2 Experiment 2: Fingertip Position Tracking Accuracy

The final experiment yields the depth resolution of the system. In this experiment, a step ramp was placed underneath the middle finger (d3). Keeping the step ramp in place and the middle finger against the step ramp, the hand is pulled toward the wrist or positive X direction. The fingertip yielded a relatively constant Y value, and an X and Y position that linearly changed with hand position and ramp height, respectively. To estimate the depth, the X data was fit with a straight line. The straight line was then used as the horizontal value for the graph of Z in Fig. 10. A second line was fit to the Z coordinate and subtracted from the data in Fig. 10, leaving the noise. The slope of the line in Fig. 10 is Δz/Δx = -0.1878 and a standard deviation of 0.4971 mm.

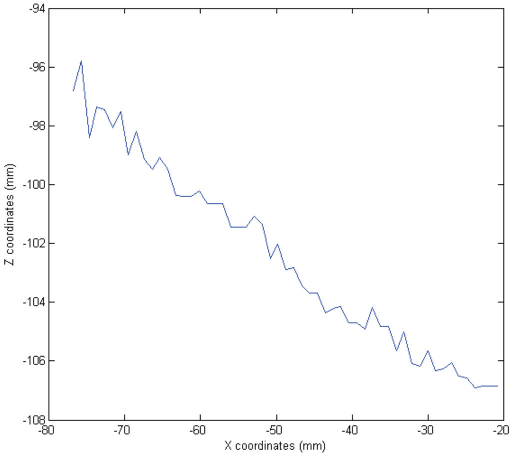

Fig. 10. Z coordinate of d3 fingertip as a function of X position on step ramp.

4.3 Discussion of Numerical Efficiency

For convenience, this research used offline processing of two different systems. That is the Lock and Hold scanner and the fingertip tracker. As such, the process would need to be combined and optimized for practical commercialization. For the acquisition component we introduced a new method called Lock and Hold SLI. In this application the Lock process is typically used to acquire a high resolution surface scan. In a non-contact human computer interface, this would not be necessary. Using only the stripe pattern shown in Fig. 6 (right) we could obtain lock by what we call leading edge lock where as the hand enters the camera FOV, the leading edge of stripes on the hand are identified and used to lock onto the hand surface. The absolute depth of the hand may be lost in this process but the relative depth of the hand and the fingertips is retained. Thus, only a single slide projection is necessary. In our experiments we capture 1.5 megapixels of data and then after initial preprocessing to 3D coordinates the result is downsampled by a factor of 150 to about 10,000 points. This takes about 1 second per frame using a dual core 2Ghz Intel Centrino processor. In a production system, this downsampling could be done upfront without preprocessing, with a lower resolution camera such as a 640 x 480 pixel camera. The processing is linearly proportional to the number of stripes and pixels used along the stripes. In theory we could have a 150x improvement but from experience we would expect at least a 15x improvement in speed primarily limited by the initial downsampling which involves an averaging process. The fingertip detection process runs at about 10 frames per second and uses a global correlation. Once the fingertips are located, the method could be adapted to local partition tracking (Su and Hassebrook, 2006) so if there are 5 partitions each of 1/25 the area of the entire scene, then the net speed up would be at least 5x and the partition filters could be optimized for each fingertip thereby achieving more robust and accurate tracking. So with a standard laptop Intel Centrino, we would expect to process at least 15 frames per second with just basic optimization. If a GPU or imbedded processor were used then the speed up would be considerably more and we would conjecture that the system could run at the frame rate of the camera.

5. Conclusion

Human to computer interfaces have been so far dominated by hand held and/or physical interfaces such as keyboards, mice, joysticks, touch screens, light pens, etc.. There has been considerable study in the use of non-contact interface technology that use image motion, stereo vision, and time of flight ranging devices. Using image processing of a single camera image, there is difficulty segmenting the feature of interest and poor depth accuracy. Stereo vision requires two cameras and is dependent on distinct features on the surface/object being measured, and time of flight systems are very expensive and lack close range accuracy.

We believe that Structured Light Illumination is a practical solution to the non-contact interface problem because of the simplicity of one camera and one projector, and its direct and accurate measurement of human hands and faces. Furthermore, with the advent of projected keyboards for augmented reality interfacing, a camera and projector are already present. In fact, the keyboard pattern could be used as the SLI pattern. In general, SLI, particularly the single pattern methods described in this research, are accurate, surface feature independent, and require only a simple slide projection in either visible or Near-Infra-Red light frequencies. The illumination source only requires efficient LED based illumination technology. As discussed in the results section, the accuracy of the depth measurement is within 1 mm so the demonstration is not just a non-contact "mouse" but a five finger analog controller. Full finger motion control could be used for a wide range of augmented reality interfacing that could be as simple as mouse and keyboard control or as sophisticated as a musical instrument interface or possibly even a sign language interface.

6. References

M. Betke J. Gips, and P. Fleming, "The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access For People with Severe Disabilities." IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10:1, pp. 1-10, (March 2002).

C. Guan, L.G. Hassebrook, D.L. Lau, "Composite structured light pattern for three-dimensional video," *Optics Express,* 11(5): pp. 406–17 (2003).

O. Hall-Holt and S. Rusinkiewicz, "Stripe Boundary Codes for Real-Time Structured-Light Range Scanning of Moving Objects," *Proc. Int'l Conf. Computer Vision*, pp. 359-366, (2001).

L.G. Hassebrook, D.L. Lau, "Structured Light Illumination Strategy INTRODUCTION OF LOCK AND HOLD STRUCTURED LIGHT ILLUMINATION," University of Kentucky EE Report #CSP 06-004, 2-20-06, Revised 5-26-06

Jielin Li, Laurence G. Hassebrook, and Chun Guan, "Optimized two-frequency phase-measuring profilometry light-sensor temporal-noise sensitivity," *J. Opt. Soc. Am. A*, 20(1), (2003).

M. Maruyama and S. Abe, "Range sensing by projecting multiple slits with random cuts," *IEEE Trans. Pattern. Anal. Mach. Intell.* 15, 647–651 (1993).

H. Morita, K. Yajima, S. Sakata, Reconstruction of surfaces of 3-d objects by *m*-array pattern projection method, in: IEEE International Conference on Computer Vision, pp. 468–473 (1988).

Jaihui Pan, Peisen S. Huang, and Fu-Pen Chiang, "Color-phase shifting technique for three-dimensional shape measurement," *Optical Engineering* – Vol. 45, Issue 1, 013602, (January 2006)

M.A. Rodrigues, A. Robinson, W. Brink, "Issues in Fast 3D Reconstruction from Video Sequences", Lecture Notes in Signal Science, Internet and Education, Proceedings of 7th WSEAS International Conference on MULTIMEDIA, INTERNET & VIDEO TECHNOLOGIES (MIV '07), Beijing, China, pp 213-218, September 15-17 (2007).

G. Schmaltz of Schmaltz Brothers Laboratories, "A method for presenting the profile curves of rough surfaces," Naturwiss 18, 315–316 (1932).

Wei Su and L. G. Hassebrook, "Pose and position tracking with Super Image Vector Inner Products" *Applied Optics*, Vol. 45, No. 31, pp. 8083-8091 (November 2006).

Song Zhang and Peisen S. Huang, "High-resolution Real-time 3-D Shape Measurement," *Opt. Eng.*, Vol. 45, No 12 (2006).

Human Computer Interaction: New Developments
Edited by Kikuo Asai

ISBN 978-953-7619-14-5
Hard cover, 382 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821