3D Imaging for Hand Gesture Recognition: Exploring The Software-Hardware Interaction of Current Technologies

Frol Periverzov and Horea T. Ilieş*
Department of Mechanical Engineering
University of Connecticut

January 8, 2012

Abstract

Interaction with 3D information is one of the fundamental and most familiar tasks in virtually all areas of engineering and science. Several recent technological advances pave the way for developing hand gesture recognition capabilities available to all, which will lead to more intuitive and efficient 3D user interfaces (3DUI). These developments can unlock new levels of expression and productivity in all activities concerned with the creation and manipulation of virtual 3D shapes and, specifically, in engineering design.

Building fully automated systems for tracking and interpreting hand gestures requires robust and efficient 3D imaging techniques as well as potent shape classifiers. We survey and explore current and emerging 3D imaging technologies, and focus, in particular, on those that can be used to build interfaces between the users' hands and the machine. The purpose of this paper is to categorize and highlight the relevant differences between these existing 3D imaging approaches in terms of the nature of the information provided, output data format, as well as the specific conditions under which these approaches yield reliable data. We also explore the impact of each of these approaches on the computational cost and reliability of the required image processing algorithms. We highlight the main challenges and opportunities in developing natural user interfaces based on hand gestures, and conclude with some promising directions for future research.

Keywords: 3D Sensing, Structured light, Time-of-Flight, Hand Gestures, Virtual Reality, Survey.

1 Introduction

The ever increasing amount of synthetic 3D information that we need to interact with every day is driving the need for more efficient means to explore and manipulate spatial data. Major software giants, such as Apple, Google and Microsoft, have recently patented their concepts of a 3D desktop interface for their operating systems, and the gaming industry appears to have the lead in commercializing novel 3DUIs for interaction with spatial data. The common agreement seems to be that human gestures form a powerful paradigm for building more intuitive 3D user interfaces for making sense of and manipulating synthetic spatial information.

This, of course, is not at all surprising. We start 'waving' our hands and use hand gestures to interact with our 3D environment before we can speak. Later on, we use hand gestures when we

*Corresponding author.

tell stories, or provide nuances or deeply embedded categories of meaning (Figure 1), and there is data showing that hand gestures are closely related to our spatial perception and visualization [1]. Consequently, one can conjecture that one of the most promising 3DUI paradigms for manipulating 3D information is one in which the user interacts with the spatial data with his/her bare hands, i.e., without the need of wearable hardware.

There is a large amount of effort being spent on developing hand gesture-based 3DUIs for manipulating 3D information, and the available technologies are evolving rapidly, which suggests that a review of the state of the art of the relevant 3D imaging technologies is timely.

Figure 1: Hand gestures have deeply embedded categories of meaning [2].

The earlier techniques used 2D images and required the user to wear wired or wireless hardware, which has been proving cumbersome and ergonomically challenging for any spatial tasks of reasonable complexity. On the other hand, the advances in computing power and computer vision hardware and software opened the door to novel 3DUIs that do not require wearable hardware, and hence, they do not restrict, in principle, the hand movement. Nevertheless, these methods can be computationally expensive, and can be influenced by many of the standard environmental factors that affect the performance of computer vision systems.

Building fully automated systems for *tracking*[1] and *recognition* of bare hand gestures requires robust and efficient 3D imaging techniques as well as potent shape classifiers. These tasks require hardware and software that must handle several key issues:

req 1: *High-speed sensing*: fast movement of the hand and motion with translational speeds up to 8 m/s and angular speeds up to 300 degrees/second [3, 4];

req 2: *Occlusion management* of the frequent finger/hand occlusions;

req 3: *Sensing resolution*: hands and fingers have a relatively small size compared to the upper and lower limbs and fingers are often clustered;

req 4: *Robustness* against sudden changes in the lighting conditions and background within the environment where the hand motion takes place;

req 5: *High-dimensionality of the models used for gesture recognition*: large number of parameter estimations due to the multiple degrees-of-freedom of the hand;

req 6: *Complex gesture semantics*: gestures can be either static, dynamic or both; are intrinsically ambiguous, and vary from person to person.

Hence, any systems designed to track and recognize hand gestures must generate and handle large amounts of data in an efficient manner in order to minimize latency, must be robust against changing environments and partial occlusions. Once the hand data is segmented, and a geometric model of the hand is available, shape classifiers must be employed to interpret gestures.

The large number of research and review papers published in the last few years on vision-based gesture recognition systems is a clear sign of the increasing interest on this topic. This paper is meant to complement the published literature by focusing on important aspects of 3DUIs that

[1] In this paper, tracking includes hand detection and segmentation.

are omitted from existing surveys. Specifically, our review focuses on the more recent 3D imaging methods used to track hands, and on the important hardware-software interaction. Its scope is to categorize and highlight the relevant differences between existing 3D imaging approaches in terms of nature of the information provided, output data format, as well as their robustness against changes in environmental conditions. This, in turn, allows us to explore the impact of each of these approaches on the computational cost and reliability of the required image processing algorithms. While gesture recognition is a key aspect of any such system, it is outside the scope of this paper. Good recent reviews focusing on the existing approaches to gesture recognition appear, for example, in [5, 6, 4].

One of the most recent reviews of the tracking and recognition methods used for hand gesture recognition is presented in [7], but without a discussion of how the different sensing methods influence the gesture recognition process. One classification of hand gestures is presented in [8], while [9] investigates issues related to multimodal interfaces using hand gesture recognition and speech. A survey of vision based human machine interfaces built for medical applications is presented in [10], and [11] presents a brief overview of several methods used in gesture recognition. Review articles that strictly concentrate on gesture recognition without treating the technical aspects of the 3D interface come in even larger numbers as discussed in section 3. General issues related to user interfaces are reviewed in [12, 13, 14] and there is a large number of papers on other potential technical approaches, such as like brain-computer interfaces [15], and multi-modal approaches [16]. Although promising, these approaches are not yet capable of addressing the hand gesture recognition problem. Consequently, these approaches are outside the scope of this paper.

Outline

We review in section 2 the current commercial solutions used for manipulating objects in 3D. Section 3 surveys and analyzes different emerging technologies that can be used to build hand-gesture driven 3DUI. We first review the working principles, followed by a detailed analysis of the relevant differences among the most competitive technologies and a summary of current performance indicators, as well as of resolution limiting factors. We conclude in section 4 with a review of the main challenges and opportunities in developing natural user interfaces based on hand gestures, and identify some promising directions for future research.

2 Commercially Available 3DUI Solutions Using Hand Gestures.

2.1 3DUIs for Hand Gestures Employing Hand Held Devices

Probably the earliest 3DUI specifically designed for manipulating 3D information is the 3D mouse, which offers up to 6 degrees of freedom (DOF) for the input. It requires the user's hand to be in contact with the input device, and 3D mice have found some popularity in industrial design and CAD communities. These devices have been designed to capture the often subtle input from the user's fingers that are required for accurate interactions with a CAD model, but the functionality is limited to standard tasks such as select, pan, zoom and rotate the model or camera. There are several variations available on the market [17, 18, 19, 20]. Moreover, the electronics and gaming industries have begun to commercialize a set of handheld devices that offer control of multiple degrees of freedom, from 3D pointing devices by Panasonic, which provides control of three degrees of freedom, to 6 DOF input devices such as Nintendo's Wii and Sony's Playstation controllers. These handheld devices are effectively more complex 3D mice, but are not designed to capture hand gestures.

A new set of controllers are those providing not just multiple DOF input, but also haptic feedback through spatial forces and even torques around the three coordinate axes. Several manufacturers provide a wide range of solutions for the personal and professional uses such as Novint [21], Sensable [22], and Haption [23]. These devices are actuated mechanisms and their workspace is limited by specific geometric and kinematic limitations of the actuated mechanisms.

Data gloves have been proposed as input devices for capturing the position and orientation of the user's hand and fingers, including the angles corresponding to the bending of fingers. There are both wired and wireless versions available [24, 25, 26, 27], but these involve wearable hardware, such as sensors, and power sources for the wireless transmitters. Furthermore, data gloves attached to a powered mechanical exoskeleton have been developed for applications requiring haptic feedback. These 3DUIs can offer a very large workspace and high data tracking resolution. The downside is the impact on usability driven by the often heavy wearable hardware, which, in turn, leads to increased fatigue in the user's arms[2] when used for a prolonged period of time.

2.2 3DUIs for Hand Gestures Without Hand Held Devices

2.2.1 3DUIs with Workspaces Smaller than the Span of the User's Arms

- *Capacitive Touch Screen:* Cypress introduced TrueTouch [28], which is a touch screen able to detect the 3D position of an object placed not farther than 5 centimeters from the screen. The sensing is performed by an array of capacitive sensors integrated into the screen. Due to the fact that the entire interface can be incorporated into the touch screen, it could be used as a 3D hand interface for portable devices such as smart phones, and tablets. However, the relatively small workspace places limitations on the usability of this interface, and the range of gestures that can be used for manipulating 3D information.

- *Ultrasonic Detectors:* Both Elipticlabs [29, 30, 31] and Nokia [32] have patented their versions of a gesture based touchless interface that can detect the position of the user's hand in a space located within 20 cm from the sensing device. Their interface is based on ultrasonic emitters that can detect the position of the hand by analyzing the acoustic wave reflected by the hand, and the distance is computed by using the standard principle of triangulation. Elipticlabs has already demonstrated the detection of waving gestures, and the movement of the hand towards or away from the ultrasound receivers.

- *GestureCube:* IDENT Technology AG is developing GestureCube [33], a commercial product in the shape of a cube that has mounted displays on 5 of its sides and incorporates several multimedia features. The sensing uses an electric field around the cube and the sensors read the disturbances in that field created by the user's hands. This user interface is able to detect hand gestures in close proximity of the cube. Although the GestureCube is not designed to be used as a 3DUI for manipulating objects in 3D, it was demonstrated that it can detect translational and rotational hand movements. This interface can be used to detect some finger movement as well as long as the hand is near the displays.

- *The Bi Directional (BiDi) Screen* [34] is currently being developed by MIT Media Labs and relies on one of the technologies currently used in multitouch LCD screens. The detection of multiple simultaneous touching events occurs by using an array of light sensors integrated into the pixel array of the screen. The depth of view of this sensor array has been recently increased, which allows the sensors to detect the depth map of an object located in front

[2]This increased fatigue in the user's arms is known as the 'Gorilla Arm' effect.

of the screen at a distance smaller than approximately 50 cm. A commercial product with similar characteristics is built by Evoluce [35], but the company does not provide the details of the sensing technology being employed. Nevertheless, is has been demonstrated that by using this 3DUI one can detected gestures like waving left-right, up-down as well as and and the movement of the hand towards or away from the screen. Due to the relatively larger sensing range, this technology might enable the detection of more elaborate gestures.

2.2.2 3DUIs with Larger Workspaces

Marker-Based Tracking Systems: Oblong Industries has released the g-speak^{TM}[36] platform that can be used to detect the user's hand gestures in a virtually unlimited space by tracking passive markers. The position of these markers can be determined in space by performing photogrammetry of data streamed from 2 or more video cameras. Oblong Industries implements this method by incorporating the reflective markers into a regular thin glove which the user needs to wear while manipulating virtual objects. This system can track finger gestures in a space limited, in principle, only by the location and number of the video cameras used.

2.2.3 Specialized Software Development Kits (SDK)

Several SDKs that have been launched over the past two years can be used, in conjunction with appropriate 3D imaging techniques, to track the motion of the human body and provide the kinematic parameters through an articulated kinematic model of the body. Importantly, these methods do not require wearable hardware to perform the tracking. There are four such SDK released so far, namely iisu^{TM}by Softkinetic [37], Kinect for Windows by Microsoft [38], Maestro3D by Gesturetek [39], Bekon by Omek [40] and the OpenNI framework [41]. Microsoft Research has recently introduced KinectFusion [42], which uses the depth data from a handheld Kinect sensor to track the sensor and reconstruct the physical scene in real time with a GPU implementation. Moreover, several other companies, such as Mgestyk [43] are working on developing their own SDKs. Hence, one can expect a strong pace of software development and increased competition over the next few years.

The existing SDKs support multiple 3D cameras commercialized by several companies, such as PrimeSense [44], Baumer [45], Canesta, MESA Imaging [46], PMD Tec [47], Panasonic [48], Optex [49], SoftKinetic [50] and Fotonic [51]. The 3D cameras commercialized by these companies use the Time Of Flight (TOF) or structured light (SL) 3D imaging principle, which are only two of the several existing 3D imaging techniques. These and other methods are discussed in section 3.

In principle, systems built with TOF/SL 3D cameras and the SDKs mentioned above used as middleware should be capable to detect hand and finger gestures without requiring wearable hardware in a workspace that is sufficiently large for most practical applications. Nevertheless, the performance characteristics achievable by these cameras are not sufficient to reliably and robustly detect finger movement. These aspects as well as potential alternatives are discussed in detail in section 3.

3 State of the Art 3D Imaging Techniques

Three dimensional imaging of physical objects has been a very active area of research for over two decades and is going through a period of heightened attention due to multiple advances in imaging hardware, software and computing power. The basic task of 3D imaging methods is to construct a

geometric model of a physical scene being observed, usually in terms of a point cloud that is later post-processed.

In the broadest sense, the existing approaches to 3D imaging either require physical contact with the object, such as the coordinate measuring methods, or compute geometric properties from data collected by non-contact sensors as summarized in Figure 2. Since we focus here on 3D imaging for hand gesture recognition, we do not discuss the approaches that exploit the transmissive properties of the objects being observed, or those that rely on non-optical reflective properties of the object surface. Furthermore, the fact that the user is in the illuminated scene places strict constraints on the safety of the illumination sources as discussed below.

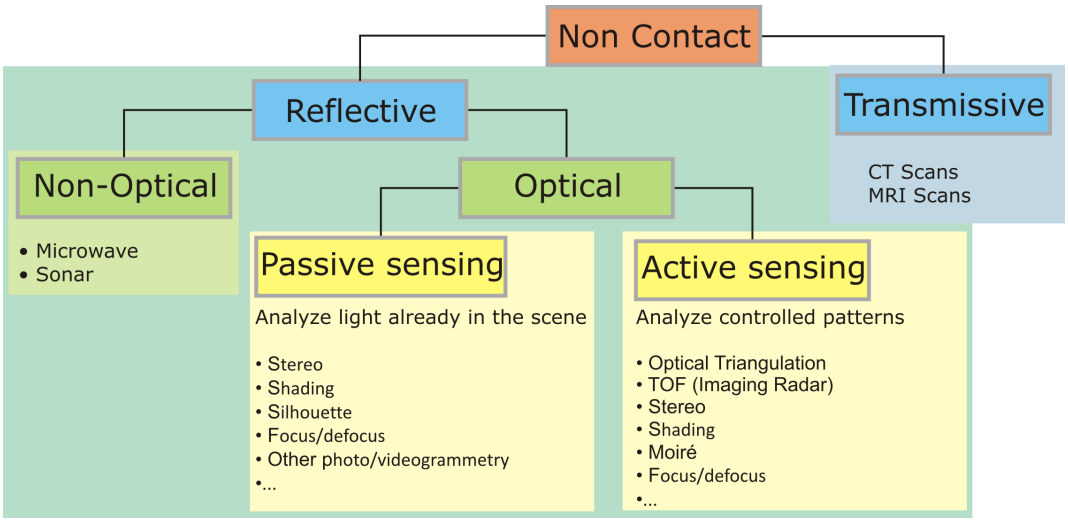

Figure 2: One taxonomy of 3D imaging approaches after [52].

All optical 3D imaging methods analyze interactions between electromagnetic radiation and the scene under observation. The passive approaches exploit the existing illumination in the scene being investigated, and tend to work 'well' under near-ideal illumination conditions. These methods look for visual cues in the images or sequence of images that they operate on to extract the geometric properties of the object. On the other hand, the active imaging methods project an electromagnetic wave, typically in visible or infrared spectra, onto the scene, and measure the changes in specific properties of the reflected waves, which are then mapped to geometric quantities. As a general rule, the performance of all optical imaging methods depends, although not to the same extent, on the illumination conditions, and specific surface properties of objects, such as differential properties, reflectance, and opacity, as well as specific hardware capabilities. A detailed historical overview of the active methods can be found in [53].

For the rest of this section, we present the working principles of those methods that appear to be the most promising for our task of detecting, tracking and recognizing hand gestures. We also analyze the relevant differences among existing approaches with respect to factors affecting the resolution, reliability, and computational cost of the algorithms that are needed to process the raw data output by the accompanying sensors.

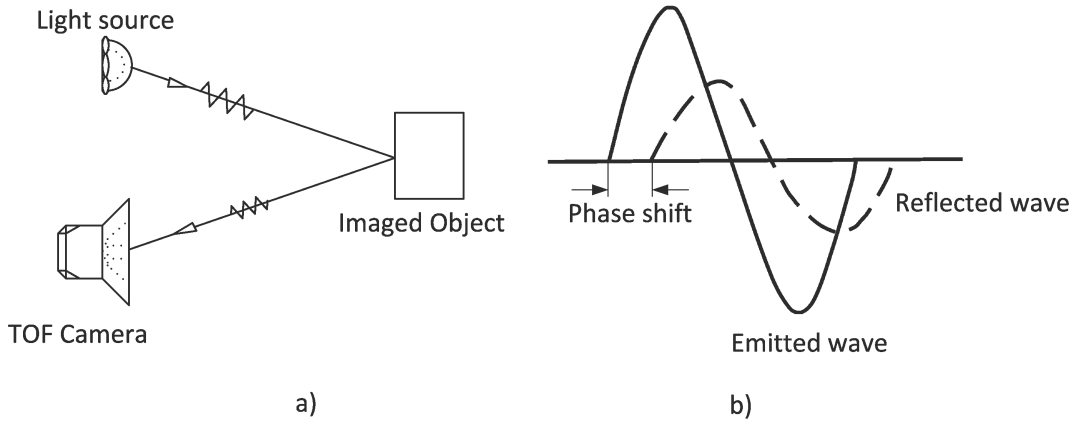

Figure 3: TOF range measurement principle: a) Typical hardware setup; b) Phase shift measurement

3.1 Measurement principles

3.1.1 Time of Flight (TOF) 3D Imaging

The TOF functioning principle relies on measuring the time that a controlled electromagnetic wave travels from the source to the sensing device, which captures the wave reflected by the target object [54] as illustrated in Figure 3. Since the speed of light is constant, the measured time can be easily transformed into a distance value. The measurement procedure goes through the following steps: a modulated wave is emitted from the source; the wave reflected by the object and captured by the sensing device has the same frequency as the emitted wave, but a smaller amplitude and a different phase [55]; the phase shift between the emitted and captured waves, which is proportional to the travel time of the wave, is then mapped to a distance value.

The TOF cameras capture entire scenes with each laser or light pulse rather than performing sequential scanning. The time of flight is measured from either phase differences between modulated emitted and imaged pulses captured by CMOS or CCD sensors, or a dedicated shutter system [56, 57]. These TOF cameras output evenly distributed range and intensity images, and avoid the correspondence problems of stereo vision or structured light systems as discussed below and in [54]. Furthermore, the surface reflectance of the objects in the scene has a much smaller influence on the range data output by the TOF cameras than for other optical systems. However, the delay needs to be measured with very high accuracy, and the depth measurement resolution depends on the modulation frequency and the measurement's non ambiguity range (NAR)

$$\Delta R = NAR \cdot \frac{\Delta\varphi}{360^o} \tag{1}$$

where φ is the measured phase shift [54]. For example, current sensors can achieve a resolution of 1 cm for a working range of 0.3-7m [47]. Increasing the depth resolution to 5mm requires a reduction in the camera's working range by a factor of 2.

3.1.2 Optical Triangulation with Laser and Structured Light

Triangulation is one of the fundamental principles used by a number of range sensing techniques, and laser triangulation is one its most common applications. Lasers are compact, and offer great

control of both wavelength and focus at large distances. There are many embodiments of this principle that differ in the type and structure of the illuminating source, and of the sensor.

The measurement procedure, as described for example in [58], uses a laser module as a light source and a regular video camera that are set up as shown in Figure 4. All geometric parameters that are known at the start of the measurement procedure are shown in blue, including the relative position and orientation of the laser module with respect to the camera. The laser module projects a planar laser 'sheet' onto the inspected object. Point S' represents the image of point S as 'seen' by the camera, and, consequently, its image plane coordinates are known. By determining the angles $\angle SFL$ and $\angle FLS$, and by observing that distance LF is known, one can determine the spatial coordinates of S from the LSF triangle. The coordinates of the visible boundary points are obtained by repeating this process while sweeping the object with the laser plane. Some of the typical concerns of these popular methods are the time needed to mechanically sweep the scene, eye safety when dealing with human users, as well as noise and depth resolution.

Rather than projecting a laser plane onto the object, one can project a 2D pattern onto the inspected surfaces, or so called 'structured light', and use measured changes in the reflected pattern to compute distances [59]. Some of the commonly used patterns are fringes [60], square grids [61] and dots [62], while the wavelength of the structured light can be inside or outside the visible spectrum. These methods need to assign each element of the captured light pattern to the correct element of the emitted light pattern as illustrated in Figure 5. Establishing this correspondence is one of the major challenges of this technique, and several pattern coding strategies have been proposed. These solutions strongly affect the performance of the structured light imaging technique as discussed below. A good discussion of all the relevant steps involved in building a structured light 3D scanner can be found in [63].

3.1.3 Stereo Vision

The typical stereo technique uses 2 cameras whose relative position is known, and looks for the differences in the images observed by the cameras in a manner similar with how human vision functions. The stereo vision can either be: (a) passive, case in which the features observed by cameras under natural illumination conditions are matched, or (2) active, by projecting artificial texture on the scene for improving the feature matching. Both approaches use the triangulation principle to produce the depth map.

The construction of the depth map depends on matching the image points captured by the cameras to the corresponding physical point, and uses the triangulation principle. As shown in Figure 6, the image planes may be aligned[3], and the depth information Z is obtained from two similar triangles SS'S" and SO_1O_2. For a 2-camera passive setup, the accuracy of the depth measurement decreases according to a quadradic law:

$$\partial Z = \frac{-Z^2}{f \cdot b} m \tag{2}$$

where **f** is the focal length of the lenses mounted on the two cameras; **b** is the stereo baseline; and **m** is the correlation accuracy that depends on the specific resolution [65].

The passive (dual or multi-view) stereo systems are using triangulation algorithms that require feature matching among the various viewpoints, which is an open problem [66]. Realtime stereo systems are emerging [67], but difficulties in matching features continue to influence their robustness. A more robust feature matching process is obtained by using active illumination. However,

[3]Note that non-aligned image planes is sometimes known as photogrammetry [64].

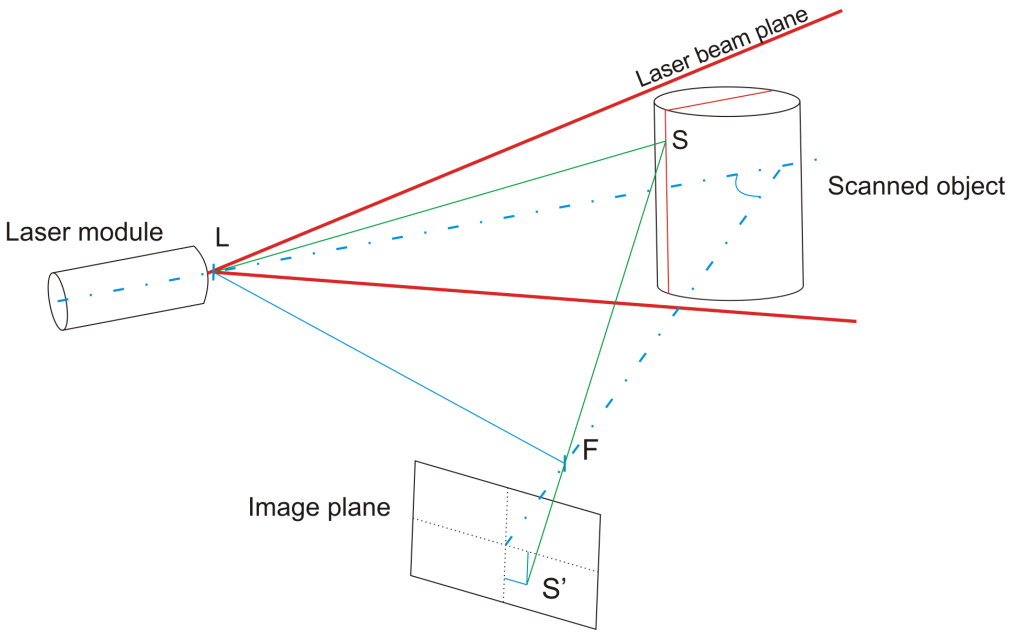

Figure 4: Principle of laser triangulation

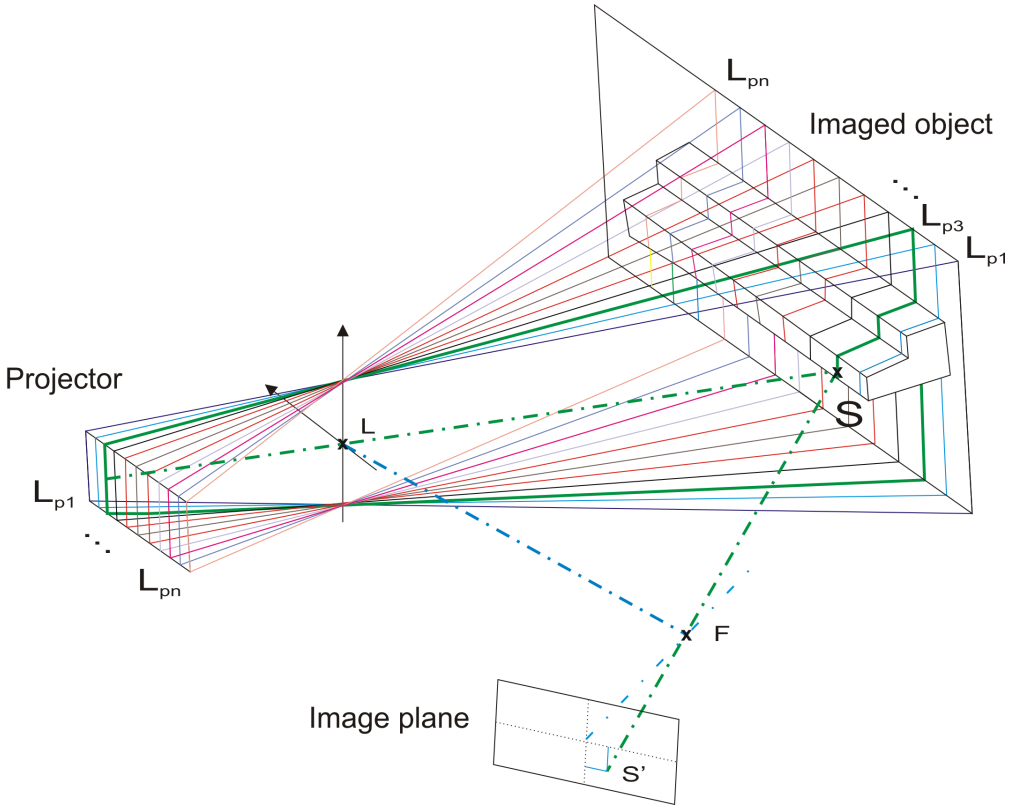

Figure 5: Structured light 3D imaging with fringe patterns.

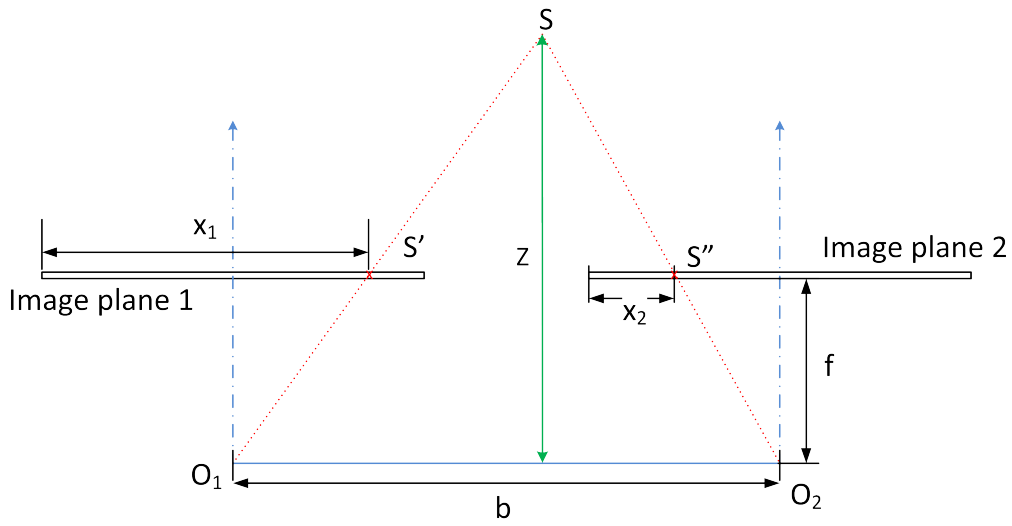

Figure 6: Triangulation for a 2-camera passive stereo setup

the two-camera active stereo vision setup suffers from potential ambiguity in solving the correspondence problem, which, in turn can lead to false matches [52]. These ambiguities can be resolved by using multiple cameras [64], also known as photogrammetry.

3.1.4 Optical Interferometry and Moiré Methods

Optical interferometry techniques project a light pattern (e.g., monochromatic, multiple wavelength, white-light) onto a scene and analyze the interference of the reflected patterns with a prescribed reference pattern. From the phase difference that occurs between the reference and the reflected patterns one can infer the depth map with a resolution on the order of nanometers [68, 69]. Digital holography can be used for inspecting large volumes by capturing only one instantaneous 2D intensity image. Methods based on digital holography can achieve an image resolution in the range of micrometers but the reconstruction process of the 3D scene requires seconds for each 3D frame [70, 71]. Furthermore, large scenes require lasers to generate coherent light, which, in turn, generate speckles, and problems with phase ambiguities for surfaces with sharp discontinuities [72, 73]. Moreover, lasers raise safety concerns when human users are present. For overcoming these limitations, one can use optical coherence tomography [74] that results in depth maps of micrometer resolution. However, coherence tomography can only be used for depth ranges on the order of centimeters [75, 72]. Note that optical interferometry methods require high intensity light sources, and highly stable opto-mechanical setups [76].

Moiré techniques illuminate the scene with a periodic light pattern and capture the image as seen through a high-frequency periodic grating whose orientation is prescribed [77, 78]. The geometric information is extracted from analyzing the interference in these patterns, which give accurate descriptions of changes in depth [79]. These methods have an accuracy of up to 10 microns. Ambiguities in measuring adjacent contours is typically resolved by taking multiple moiré images with repositioned gratings.

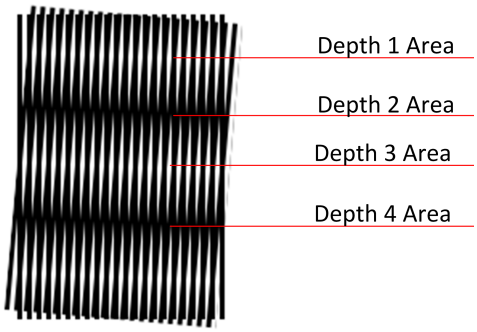

Figure 7: Moiré patterns

3.1.5 Fresnel Holograms

Typical holography relies on the coherence properties of lasers. However, several holography techniques that use incoherent (white) light instead of lasers have been proposed. A recent review of three specific techniques using diffraction of incoherent light through a set of concentric rings known as Fresnel Zone Plates is found in [76]. The distance information of the object is encoded in the density of the rings because the points that are closer to the system produce less dense rings than those that are farther. In other words, the Fresnel Zone Plates are parameterized with respect to the distance between the imaged surfaces and the projector. Construction of multicolor 3D holograms acquired in real time has been shown in [80].

3.1.6 Shape from Shading

These methods extract accurate 3D shape information from an imaged scene (one or more images) by using the shading cues detected in the images under controlled illuminating conditions. Recovering the geometric information requires known surface reflectance of the objects in the scene, a constrained reflectance map, or multiple images and light sources [81]. The method uses the intensity variation at each pixel to estimate the normal of the surface at the surface point that projects to the image pixel. Shape from shading with one light source is ill-posed in general with no unique solution [82]. On the other hand, by illuminating a surface with at least two linearly independent *known* light sources, a unique depth map can be recovered. More recent robust reconstructions from shading cues have been achieved by using multiple images with changing illumination taken from the same view-point, or by using linearly independent colored-light stereo sources whose geometric location and orientation are known [81].

3.1.7 3D Integral Imaging

This technique is based on the principle of integral photography [83, 84] and the 3D image is obtained by processing multiple images taken from coplanar, grid-aligned imaging sensors or lenses. The depth information is generated by analyzing the relative shift of the position of an object in these different images [85]. The reconstruction of the 3D image can be performed by back projecting the rays that have generated the different captured images. These images that are projected back will overlap, and will form a sharp image only at the Z distance at which the inspected object is located. The Z value is controlled by the distance between the images representing different viewing

perspectives that are back projected. A different approach is proposed by [86] where the range information is calculated only for the central pixels of these images followed by a refinement of the grid formed by these points.

3.1.8 Shape From Focus/Defocus

As the name implies, this class of methods detect points of the scene relative to the focal plane of a camera [87, 88].

For the shape from *focus* methods, the position and orientation of the focal plane relative to the camera are fixed and known, and the sharpest image regions are identified for specific focus settings by applying specific differential functions such as 3D gradient [89] and Laplacian [90]. It is intuitive that in order to measure the boundary of an object, we must 'scan' the object with the focal plane. The resulting 3D imaging method can achieve micrometer resolution [91], but the 3D reconstruction process is relatively slow (on the order of minutes for each frame processed [87]).

Shape from *defocus* techniques extract the depth information by taking two images, with two focal distances followed by an analysis of the blur differences in these images. The measurement principle uses the fact that objects located at different distances from the camera are blurred by different amounts. Blurring is typically modeled through the diffusion equation, and the 3D scene is reconstructed by solving the inverse diffusion problem [92].

Focus/defocus based methods heavily depend on the mechanical adjustment of the focal settings of the system, which severely influences the ability to perform real time 3D image acquisition.

3.1.9 Shape from Texture

The last class of methods that we survey here analyzes the deformation of individual texture elements, whose size and shape are constrained, and convert these deformations to a distance map. These methods require an a priori knowledge of specific properties of the textures of the objects, such as: the shape of the texels, the homogeneity [93], isotropy [94], spatial frequency [95], smoothness [96], and the planarity and the state of motion [97]. However, these methods are not applicable for objects whose (existing or projected) texture is not regular.

3.2 Comparative Analysis of 3D Imaging Alternatives

Despite the tremendous advances in 3D imaging hardware, there are no accepted standards and protocols to measure the performance of these systems. As a consequence, NIST is currently developing a 3D imaging performance evaluation facility along with the protocols for characterizing the performance of various imaging systems [98]. The NIST efforts are focusing on analyzing the effect of several factors on the performance of the systems, namely range, angle-of-incidence, reflectivity, azimuth angle, single vs multiple point measurements, and type of imaged object. We note that in the absence of such a standard, the direct comparison of the depth and image resolution, as well as sensitivity to optical parameters of the 3D imaging systems that are published by the developers or manufacturers can be misleading.

We review here two important aspects that significantly affect the performance of several key imaging techniques, and present a summary of current published performance indicators for several promising 3D imaging approaches.

Point Matching

All imaging methods that use triangulation require the identification of identical physical points or features across multiple images in order to determine the depth of these physical points. In a broad sense, the point matching algorithms identify regions with similar prominent features that contain the point for which we search across these images. As a consequence, these methods can not be used single handedly for objects with smooth boundaries that do not have distinctive textures or features. A comparative analysis is presented in [66].

Pattern Matching for Structured Light Techniques

Many discrete or continuous coding strategies have been proposed for determining the correspondence between the source of the pattern projected onto the imaged scene and the reflected pattern since this task can have a dramatic implication on the performance of the imaging method [99]. Most strategies use large areas of the captured pattern to be able to compute the depth of a single point. By contrast, the method proposed in [61] uses only the connectivity of adjacent points of a projected grid pattern, but with a lower measurement resolution than is what is presented in [60]. The latter work uses sinusoidal fringe patterns whose density controls the achievable depth resolution. The 3D imaging technique presented in [60] achieves 30FPS (frames per second) for 300k points per frame. According to [100], the standard fringe projection methods cannot measure multiple objects separated in space. Instead, the same work proposes a structured light imaging method based on statistical speckle patterns, that achieves reconstruction speeds of 17.25 FPS. A more detailed presentation of other alternatives is presented in [99].

Performance of Current Implementations

Structured light techniques produce high-accuracy depth maps, but achieving real-time implementations must avoid sequential scanning of the objects. These techniques require an efficient implementation of the point matching problem, although the sensitivity of this correspondence problem to illumination and geometric parameters influence its robustness. Moreover, structured light techniques can output direct measurements only at the matched points - all others require interpolation or some local approximations, and their efficiency decreases with the increase in accuracy due to the relatively high computational overhead. In principle, the efficiency of these methods is limited only by the available computational power. On the other hand, TOF cameras are monocular, have a relatively dense depth information and constant resolution as well as high frame rates. They have superior robustness to illumination changes, and the phase shift measurement and its mapping to distance values are straightforward and computed in the camera, which minimizes the additional processing required [101]. The depth map output by TOF cameras are largely independent from textures in the scene [102]. The current resolutions achieved by TOF cameras are lower than those using structured light, as described below. The efficiency of the depth map construction by these cameras is physically limited by the photon shot noise and modulation frequency [103, 104].

One of the fastest and most robust TOF cameras [47] offers 40 FPS at an image resolution of 200x200 pixels, and the depth map is constructed in the camera, without requiring additional computational cost. Higher frame rates of up to 80 FPS can be achieved as the resolution decreases. The typical depth measurement range for TOF cameras is $[0.3, 7]$ meters, with a precision (repeatability) smaller than 3mm and a depth resolution of about 10 mm. On the other hand, structured light imaging methods based on a speckle light pattern that follows a prescribed statistical distribution appear to be one of the most robust methods in this imaging class, and frame rates of 308 FPS with a measurement accuracy of 50 μm have been documented in [105]. The

fast image acquisition method presented in [105] uses two 4630FPS high speed cameras and an acusto-optical laser deflector to generate the statistical speckle pattern and complete the matching problem by analyzing multiple captured images at what seems to be a significant computational cost. Low cost commercial cameras using structured light achieve depth frame rates of up to 30 FPS for a depth image resolution of 320×240, a usable range of $[1.2, 3.5]$ meters and a depth measurement resolution lower than that of the TOF cameras. These parameters, however, are driven by current cost constraints imposed by developers of commercial structured light cameras rather than technological limitations.

3.2.1 Major Limiting Factors Affecting System Performance

Most of the available optical imaging techniques are affected by the specific illumination conditions, and object properties affecting the reflectance. Furthermore, several imaging methods assume the relative position and orientation between the video cameras (VC) and other devices used in the imaging process to be known or computable (column 2 of table 1). At the same time, a number of active light imaging techniques require no variations in the illumination conditions (column 3 of table 1).

Method	Relative position/orientation constraints	Illumination constraints	Other minimum working conditions	Major limiting factors
Time of Flight			– Scene at least 0.3 m away from the camera [45]	– Resolution limited to about 1cm for objects placed between [0.3,7] meters. – External light with the same characteristics as the active light [54].
Structured light	x		– Location of light source determined for each measured point [60, 58]	– External light with the same characteristics as the active light [106]
Laser triangulation	x		– Imaged surfaces must be scanned. – Laser beam reflected as scattered light [64]. – Beam reflection identified into the captured image [106].	– Objects with sharp edges [64].
Passive Stereo	x		– Require point matching. Hence, smooth surfaces require projection of artificial texture or active illumination from multiple sources. – Measured points must be visible from at least 2 different locations (no occlusion) [64]; imaged object in proximity of the camera (e.g., for passive stereo vision the measurement accuracy decreases according to the quadratic law (2) [107];	– Camera occlusions [108]. – Objects must have distinctive features
Optical interferometry	x	x	– Environment factors affecting the light path (e.g., hot air, dense water vapors) [71]; non-interferometric background; ambiguity of reconstruction (solution to "twin-image" problem proposed in [109]).	– Mechanical vibrations – Lasers or high intensity white light sources can induce retinal damage [76]. – Robust pattern coding strategies
Fresnel holograms	x	x	– The wavelength of the projected light must be known [80].	
Moiré patterns		x	– The projected fringes need to be visible in the captured images [79].	– When data gathered from multiple images, the methods are sensitive to motion, blur and step discontinuities in the object boundaries [110].
Photometric stereo and shape from shading	x	x	– The points that are measured need to be illuminated by at least 2 light sources without shadows [81].	– Complex geometry producing shadows [81]. – Existing texture, ambiguity due to non-uniqueness.
Integral imaging	x		– Requires point matching [85, 86].	– When using synthetic aperture, relative location of cameras affect the imaging accuracy [85].
Shape from focus			– Adjustment of discrete focal settings [87]. – Object must have sharp edges that differentiate object from blured background [89].	– Discrete focal settings [87].
Shape from texture			– Texture must be known prior to measurement [96]. – Optical model of the camera must be known [95]. – Visible texture cues for all the reconstructed surfaces.	– Unexpected or non-measurable texture properties – Shadow or light patterns generated by uncontrolled sources.

Table 1: Major Limiting Factors Affecting System Performance

3.2.2 Resolution Limiting Factors

All methods presented have the in-plane measurement resolution limited by the resolution and the field of view (FOV) of the cameras. Moreover, all methods using multiple cameras or cameras and sensors that have prescribed positions and orientations in space are sensitive to the accuracy of their spatial location. The computationally intensive 3D reconstruction algorithms can be implemented on GPU cards to achieve at least the performance mentioned in the following table. This will, in turn, make the CPU available for the other processes. Unfortunately, most of the research articles in this field present their speed performance indicators in terms of FPS of the reconstructed 3D scene, or the ambiguous "real-time" attribute. We observe that these parameters heavily depend on specific hardware and software decisions used in each specific case[4], but these details are not provided in the literature. A summary of the major resolution limiting factors and of current performance indicators is presented in table 2.

[4]See also NIST's efforts for standardizing the measurement protocols [98].

Method	Resolution limiting factors	Current Resolutions
Time of Flight	Unavailability of low-cost, high-power IR-LEDs to produce higher modulation frequencies [54, 103, 104]	Highest depth resolution about 1 cm; largest sensor resolution is 200x200 pixels [47]
Structured Light	Global illumination [111]. Also, coding used for projected fringes affects the capability and resolution of these methods. In general, depth resolution is a fraction of the density of the projected fringes [112, 60]. Moreover, simple stripes require phase unwrapping which fails near discontinuities (as is the case of objects separated in space) [100]. Statistical speckle patterns can be used instead.	Depth maps of 50 μm [105] resolution are available at a frame rate of 308FPS.
Laser Triangulation	Resolution with which the laser beam can be shifted, and by the accuracy of the relative location of the camera and laser module. Speed of this imaging method is limited by the (mechanical) speed of scanning.	Highest depth resolution is 25 μm [113]
Passive Stereo Vision and Short Range Photogrammetry	Accuracy of the point matching task (can be improved by active illumination). The measurement accuracy decreases with distance from cameras.	− Micrometer level depth resolutions for distances lower than 50 cm, and mm level resolutions for ranges between 1-10 meters. The accuracy quickly decreases with distance - e.g., equation (2)) applies to passive stereo vision. [114]. − Depth maps of lower point density and at a lower frame rates than the TOF or the structured light techniques.
Shape from Shading	− Each point must be illuminated by at least 2 sources. Moving, deformable or textured bodies are difficult to handle unless multispectral setups are used [81].	Resolutions can be achieved at sub-millimeter levels.
Moiré Methods	− Pitch of the projected fringes [79] − [79] uses only one image (rather than multiple) which speeds up the computations	[115] shows a depth measurement resolution of 15 μm while using a commercial LCM projector capable of 1024x768 projection resolution
Optical Interferometry	Coherence and wave length(s) of light source [104]. Predominantly used over small distances up to several cm [116].	The measurement resolution can achieve very high resolution (up to tens of nanometers [117]) for relatively large scenes [71].
Fresnel Holograms	Wavelength of the light used [80].	Depth resolution of 0.5 μm and 2048 x 2048 depth points density are documented in [76, 80];
Shape from Focus	Efficiency and accuracy of focus adjustment. Relatively complex depth map extraction algorithms.	600μm depth measurement resolution [87]
Integral Imaging	Point matching requires textures or distinctive features; depth resolution limited by image resolution, FOV, number of the sensors used, and number of points that can be matched in different images.	2040 x 2040 depth image at a frame rate of 15 FPS discussed in [86]
Shape from Texture	Properties of the surface texture.	Lower accuracy than other popular methods [95, 93]; simple hardware setup.

Table 2: Resolution limiting factors and depth resolutions of current methods.

3.2.3 Salient Advantages of the Methods that are Most Promising for the New 3DUI

Time of Flight Methods

- Offer one of the lowest computational cost for the acquisition of depth maps with a depth point density up to 40000 per frame and a frame rate up to 80 FPS (at a lower resolution);
- Depth map output directly by the camera;
- Have low sensitivity to external light, as sensors perceive only infrared light, and process only frequency or amplitude modulated light;
- Can image any object that reflects scattered light;
- Have no mechanical constraints and have a simple and compact setup.

Structured Light Methods

- Offer high resolution for depth measurement and in plane measurement;
- High measurement speed and good measurement reliability for indoor imaging;
- Relatively low resolution cameras are commercially available in simple and compact setup.

Laser Triangulation Methods

- Less sensitive to texture and color changes of the imaged objects as well as to external light variations than all the presented methods except for the TOF cameras.

Fresnel Holograms

- Achieve measurement resolutions similar to the interferometric holograms without having the restriction of using coherent light;
- Can be used for measurements that go form micrometric FOV to a FOV in the range of meters [76];
- Can build multi-color holograms in real time [80].

4 Hand Gesture-Based 3DUI: Challenges and Opportunities

A definition of hand gestures can be extrapolated from the usual definition of a gesture given by Meriam Webster dictionary: a gesture is 'the use of motions of the limbs or body as a means of expression'. Consequently, we define hand gestures as the use of hands as a means of expression or providing semantics.

The discussion above suggests that there are two 3D imaging technologies that have *today* the performance characteristics, sensitivity to noise as well as the compact setup, and workspace capabilities required for a hand gesture-based 3D user interface described in section 1. In this section we review the main requirements of such a 3DUI, and explore the main challenges and opportunities.

4.1 Sensing Technology

As argued above, TOF and structured light imaging methods have today competing and somewhat complementing sensing capabilities. Structured light imaging technologies can, in principle, achieve high resolution with a relatively good performance, while TOF systems are fast, have the lowest sensitivity to changes in illumination conditions and construct the whole depth map with a relatively low computational cost required for measuring the phase shifts. Nevertheless, these capabilities are expected to rapidly evolve given the increased attention to 3D interaction. Both technologies are commercially available in compact sizes and several SDKs exist that can speed up the development of custom applications (see section 2.2.3). The TOF cameras have today a cost that is about two orders of magnitude higher than that of the Kinect sensor. The performance differences of TOF and structured-light commercial cameras suggest that the TOF cameras are preferable for tracking the rapidly moving and clustered hand and fingers (reqs. 1, 3 & 4 in section 1). Furthermore, multiple Kinect cameras can be employed for occlusion minimization (req. 2, section 1) and for tracking the body and limbs. As emphasized in section 3, the shape from focus/defocus methods as well as the shape from shading and shape from texture methods do not meet today several of the requirements identified in section 1, and we do not see them as strong candidates for constructing reliable hand gesture recognition systems.

We are developing a hand gesture tracking system uses two SDKs for facilitating hand tracking, namely the iisu^{TM}by SoftKinetic [37] as well as Microsoft's Kinect SDK [38]. The latter offers access to a simplified kinematic model of the human body and provides functions that can be used to query the model for kinematic parameters. Kinect SDK requires that the user's body be visible above the knees in order to build the kinematic model of the human body. On the other hand, iisu^{TM}SDK provides similar functions for tracking body parts, but it provides increased functionality. For example, one can define a parameterized motion in the workspace and the body part that performs the predefined motion can be identified via the SDK. Both SDKs offer multi-user tracking with some occlusion management, and can be used to track motion of limbs, but neither of them can at the moment track and recognize finger gestures.

4.2 Challenges and Opportunities

Hand tracking (detection and segmentation) is a critical step for any hand gesture recognition system. Approaches based on processing 2D images need crucial access to good features, due to the richness of variations in shape, motions and textures of hand gestures. On the other hand, three-dimensional sensing capabilities provide a direct approach to the segmentation problem. Nevertheless, the 3D imaging technologies face some of the same challenges as any other imaging system. Probably the most important observation is that no 3D vision system can be completely reliable due to the incomplete (insufficient) information captured by the sensors. Typical factors limiting the amount of captured information are resolution and noise limitations, occlusions, motion blur, as well as changes in environment conditions.

The current resolutions of *commercial* TOF and structured light cameras are insufficient for tracking ngers without decreasing the distance to subjects being observed or the eld of view (FOV) of the cameras. Some cameras allow mechanical adjustments of the FOV by modifying the focal settings of the optics. In order to use an enlarged FOV, one could use higher resolution structured light systems discussed in section 3, but such systems are not commercially available; use the additional information captured by the Kinects high(er) resolution color camera, or employ additional video cameras. Furthermore, occlusions and motion blurs, particularly of the fast moving fingers, are common failure modes in all 3D imaging and are often unavoidable. The standard approach

to handle these limitations is to use multiple cameras as well as temporal and spatial coherence information.

In our system, we track the motion of the user's body by using the structured light camera and monitor 20 joints of the skeletal model using the Microsoft SDK for Kinect as illustrated in Figure 8, including the wrists, elbows, shoulders, head, hip, and knees. Moreover, we use the wrist data output by Kinect SDK to perform robust hand segmentation from the depth map output by the TOF cameras by processing the volumetric and grayscale values as illustrated in Figure 9.

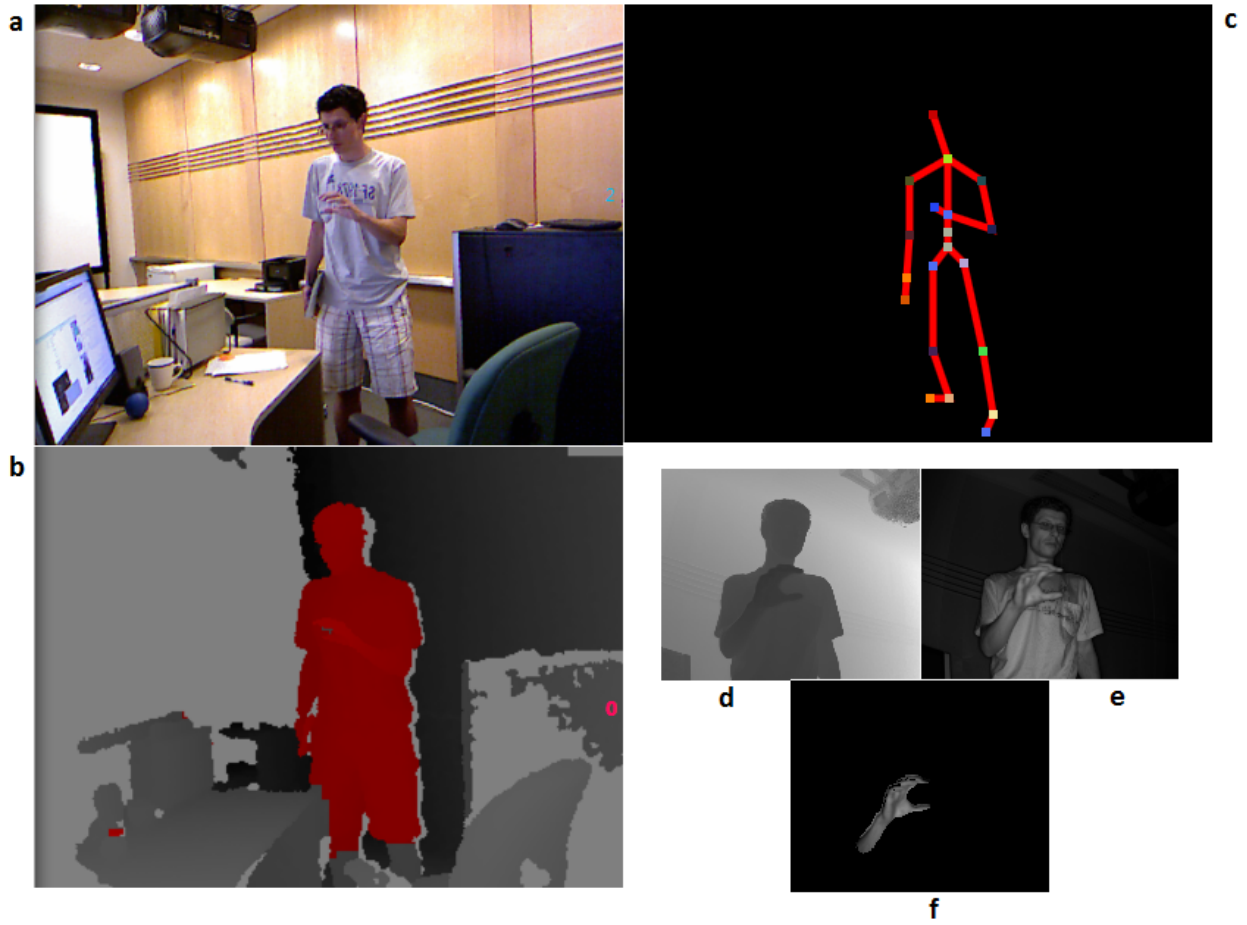

Figure 8: Hand tracking: a) Color stream data; b)The user segmented into the depth stream of the structured light camera; c) Tracked joints and the skeletal representation of the user ; d) The depth map acquired with the TOF camera; e) TOF camera intensity image; f) The segmented 3D surface of the user's hand

Hand segmentation must be followed by semantic matching of the static and dynamic gestures, which can rely on robust pose dependent shape signatures. While this step is still under development in our system, there are several different approaches that are usually employed and can be classified as either appearance based (e.g., template matching, feature extraction, skeletons) or model based (e.g., kinematic, neural networks) techniques. The existing approaches to gestures matching and interpretation are reviewed in [4, 6, 5]. A comprehensive review of gesture recognition algorithms along with an evaluation of their performance over standard data set can be found in [118, 7]. Surveys focused on hand gesture recognition methods based on vision can be found in [8, 119,

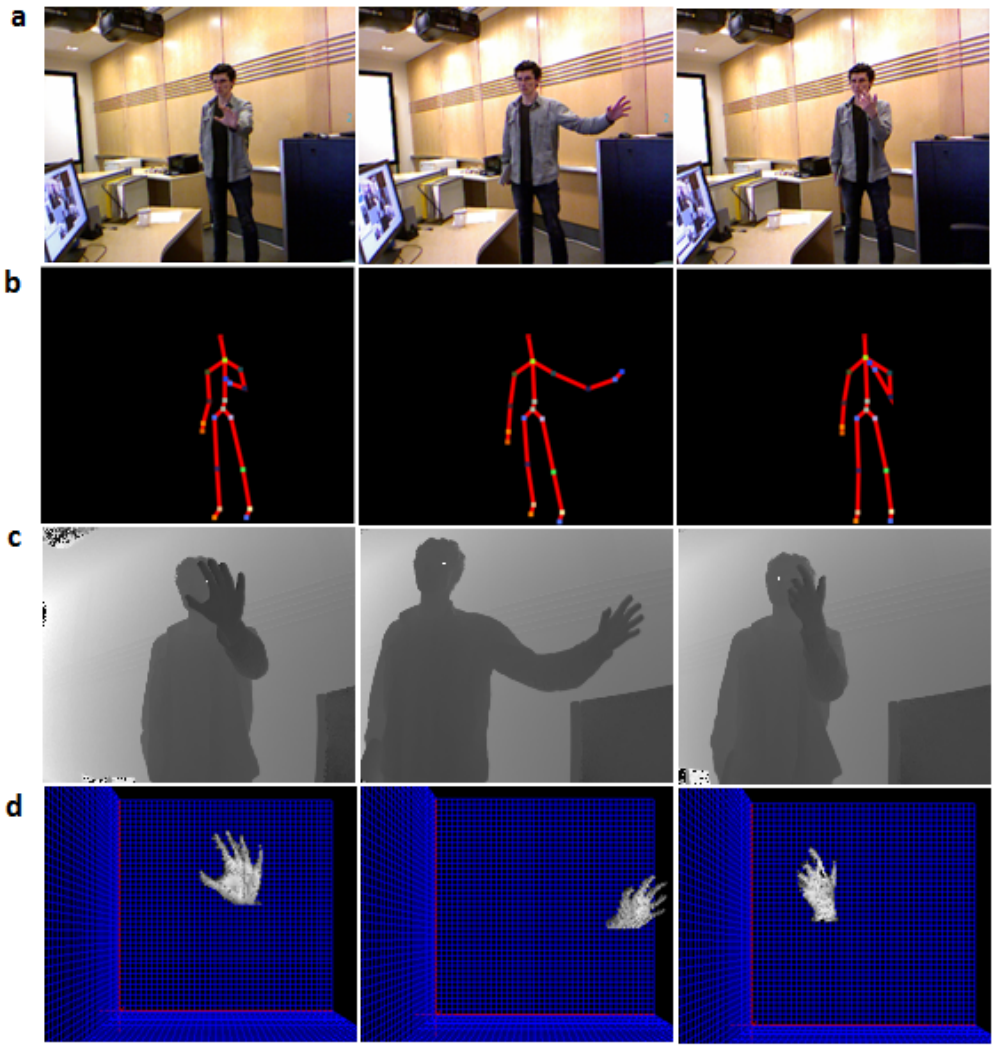

Figure 9: Hand segmentation of various hand poses: (a) Color stream data; (b) Skeletal representation; (c) TOF depth map; (d) segmented hand.

120], while a review of gesture classification methods is given in [121]. Good surveys of gesture interpretation and vocabularies can be found in [122, 123, 124].

It is important to note that defining a vocabulary of hand gestures must depend on the usability of such gestures for natural human computer interaction. There have been several usability studies of mostly 2D gestures in automotive environments [125, 126], but these results can not be applied to the problem of manipulating 3D information with 3D hand gestures. On the other hand, the current usability studies that focus on 3D interfaces [127] have limited scope and applicability given the limitations of the gesture recognition systems on which they are based. This suggests that the development of practical vocabularies for 3D hand gestures must rely on more specific usability studies that, in turn, depend on the availability of robust hand tracking and recognition technology. An exploration of the principles guiding the design of 3D user interfaces and a perspective on future research direction appears in [128].

5 Conclusions

The limitations of the existing commercial 3D imaging cameras could be overcome by integrating several depth sensing devices into one imaging system. The difficulties induced by the relatively low resolution of these commercial cameras prove to be worth pursuing, because depth information can reliably produce segmented hands in cases in which 2D image based methods may fail as illustrated in Figure 9(a). Our initial experiments show that practical low cost 3DUIs relying on natural hand-gestures can be built by combining the capabilities of commercial structured light imaging hardware with the information provided by commercial time-of-flight cameras. The availability of such a low cost 3DUI can be a 'game changer' in engineering and industrial design and provide new paradigms for the design of software and hardware interfaces, as well as for usability, technical and scientific collaboration, learning, and outreach. It is important to note that the next generation TOF sensors could quickly produce higher resolution, and low cost time of flight cameras whose costs could be comparable with the current cost of Kinect sensor [54], which would eliminate the need of a hybrid strategy such as the one discussed above.

By preventing wearable hardware attached to the user's hands, we eliminate the possibility of providing haptic feedback to the user. Nevertheless, the myriad of recent smartphone and gaming consumer applications keep proving the fact that users can rapidly adapt to environments that do not exploit the sense of touch for manipulating 3D information. This suggests that the availability of low cost 3DUIs based on hand gestures coupled with the difficulties of current haptic technologies in providing realistic haptic feedback may shift the demand for haptic systems in favor of more interactive, although haptic-less, interfaces.

Acknowledgments

This work was supported in part by the National Science Foundation grants CMMI-0555937, CAREER award CMMI-0644769, CMMI-0856401, and CNS-0927105. We would also like to thank the anonymous reviewers for their helpful comments.

References

[1] M. Chu and S. Kita, "The nature of gestures' beneficial role in spatial problem solving.," *Journal of Experimental Psychology: General*, vol. 140, no. 1, p. 102, 2011.

[2] M. Buonarroti, "Creation of Adam." Sistine Chapel, Rome, 1512.

[3] E. Varga, I. Horváth, Z. Rusák, and J. Broek, "Hand motion processing in applications: A concise survey and analysis of technologies," in *Proceedings of the 8th International Design Conference DESIGN 2004*, pp. 811–816, 2004.

[4] A. Erol, G. Bebis, M. Nicolescu, R. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, 2007.

[5] Y. Wu and T. Huang, "Vision-based gesture recognition: A review," vol. 1739, pp. 103–115, 1999.

[6] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311–324, 2007.

[7] X. Zabulis, H. Baltzakis, and A. Argyros, "Vision-based hand gesture recognition for human-computer interaction," *The Universal Access Handbook, Human Factors and Ergonomics. Lawrence Erlbaum Associates, Inc.(LEA)*, 2009.

[8] R. Hassanpour, S. Wong, and A. Shahbahrami, "Visionbased hand gesture recognition for human computer interaction: A review," in *IADIS International Conference Interfaces and Human Computer Interaction*, pp. 125–134, 2008.

[9] J. Liu and M. Kavakli, "A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games," in *2010 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1564–1569, IEEE, 2010.

[10] A. Albu, "Vision-based user interfaces for health applications: A survey," *In Proc. of Advanced Visual Computing, 2^{nd} International Symposium, Lake Tahoe, USA*, pp. 771–782, 2006.

[11] K. Shastry, M. Ravindran, M. Srikanth, and N. Lakshmikhanth, "Survey on various gesture recognition techniques for interfacing machines based on ambient intelligence," *Arxiv preprint arXiv:1012.0084*, 2010.

[12] P. Zhang, N. Li, M. Scialdone, and J. Carey, "The intellectual advancement of human-computer interaction research: A critical assessment of the MIS literature (1990-2008)," *AIS Transactions on Human-Computer Interaction*, vol. 1, no. 3, pp. 55–107, 2009.

[13] A. Agrawal, M. Boese, and S. Sarker, "A review of the HCI literature in IS: The missing links of computer-mediated communication, culture, and interaction," *AMCIS 2010 Proceedings*, 2010.

[14] N. Bernsen and L. Dybkjær, *Multimodal Usability*. Human-Computer Interaction Series, Springer-Verlag New York Inc, 2009.

[15] M. Mikhail, M. Abdel-Shahid, M. Guirguis, N. Shehad, B. Soliman, and K. El-Ayat, "BEXPLORER: Computer and communication control using EEG," *Human-Computer Interaction. Novel Interaction Methods and Techniques*, pp. 579–587, 2009.

[16] K. Kasamatsu, T. Minami, K. Izumi, and H. Jinguh, "Effective combination of haptic, auditory and visual information feedback in operation feeling," *Human-Computer Interaction. Novel Interaction Methods and Techniques*, pp. 58–65, 2009.

[17] SpaceControl GmbH, "http://www.3d-mouse-for-cad.com/," accessed September 2011.

[18] 3Dconnexion, "http://www.3dconnexion.com/," accessed September 2011.

[19] Logitech, "http://www.logitech.com/," accessed September 2011.

[20] Axsotic, "http://www.axsotic.com/," accessed September 2011.

[21] Novint, "http://home.novint.com/," accessed September 2011.

[22] Sensable, "http://www.sensable.com/," accessed September 2011.

[23] Haption, "http://www.haption.com/," accessed September 2011.

[24] 5DT, "http://www.5dt.com/," accessed September 2011.

[25] C. U. G. v2.0, "http://www.cyberglovesystems.com/," accessed September 2011.

[26] Measurand, "http://www.finger-motion-capture.com/," accessed September 2011.

[27] PixelTech, "http://www.pixeltech.fr/," accessed September 2011.

[28] Cypress, "http://www.cypress.com/," accessed September 2011.

[29] T. Dahl, "Object location." European Patent no. 2294501, March 2011.

[30] T. Dahl, "Object and movement detection." European Patent no. 2281231, February 2011.

[31] T. Dahl, G. Birkedal, and B. C. Syversrud, "Multi-range object location estimation." European Patent no. 2271951, January 2011.

[32] M. Nurmi, "User interface." US Patent no. 0256807, October 2009.

[33] GestureCube, "http://www.gesture-cube.com/," accessed September 2011.

[34] M. Hirsch, D. Lanman, H. Holtzman, and R. Raskar, "BiDi screen: A thin,depth-sensing LCD for 3D interaction using lights fields," *ACM Trans. Graph*, vol. 28, no. 5, 2009.

[35] Evoluce, "http://www.evoluce.com/," accessed September 2011.

[36] Oblong Industries, "G-speak," accessed September 2011. http://oblong.com/.

[37] SoftKinetic, "Softkinetic iisu^{TM}product datasheet v0.91," accessed September 2011. http://www.softkinetic.com/.

[38] Microsoft, "Microsoft kinect for windows sdk," accessed September 2011. http://research.microsoft.com/.

[39] G. T. Brochure, "http://www.gesturetek.com/," accessed September 2011.

[40] Omek, "http://www.omekinteractive.com/," accessed September 2011.

[41] F. OpenNI, "http://www.openni.org/," accessed November 2011.

[42] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *UIST '11*, October, 16-19 2011.

[43] Mgestyk, "http://www.mgestyk.com/," accessed September 2011.

[44] T. P. PS1080, "http://www.primesense.com/," accessed September 2011.

[45] B. T. U. G. for Digital 3D Camera, "http://ftp.elvitec.fr/," accessed September 2011.

[46] S. U. V. MESA Imaging, "http://www.mesa-imaging.ch/," accessed September 2011.

[47] P. vision CamCube 3.0 Datasheet, "http://www.pmdtec.com/," accessed September 2011.

[48] P.-I. S. Panasonic, "http://panasonic-electric-works.net/," accessed September 2011.

[49] Z. . T. C. d. Optex, "http://www.optex.co.jp/," accessed September 2011.

[50] D. D. Softkinetic, "http://www.softkinetic.com/," accessed September 2011.

[51] T. R. C. Fotonic, "http://www.fotonic.com/," accessed September 2011.

[52] B. Curless and S. Seitz, "3D photography," *Course Notes for SIGGRAPH 2000*, 2000.

[53] F. Blais, "Review of 20 years of range sensor development," *Journal of Electronic Imaging*, vol. 13, no. 1, 2004.

[54] S. Hussmann, T. Ringbeck, and B. Hagebeuker, *A performance review of 3D TOF vision systems in comparison to stereo vision systems*, pp. 103–120. November 2008.

[55] A. Nüchter, *3D robotic mapping: the simultaneous localization and mapping problem with six degrees of freedom*, vol. 52 of *Springer Tracts in Advanced Robotics*. Springer Verlag, 2009.

[56] G. Yahav, G. Iddan, and D. Mandelboum, "3D imaging camera for gaming application," in *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, pp. 1–2, IEEE, 2007.

[57] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight cameras in computer graphics," in *Computer Graphics Forum*, vol. 29, pp. 141–159, Wiley Online Library, 2010.

[58] G. Sansoni, M. Trebeschi, and F. Docchio, "State-of-the-art and applications of 3D imaging sensors in industry, cultural heritage, medicine, and criminal investigation," *Sensors*, vol. 9, no. 1, pp. 568–601, 2009.

[59] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[60] N. Karpinsky and S. Zhang, "High-resolution, real-time 3D imaging with fringe analysis," *Journal of Real-Time Image Processing*, pp. 1–12, 2010.

[61] H. Hiroshi Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi, "Dynamic scene shape reconstruction using a single structured light pattern," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pp. 1–8, IEEE, 2008.

[62] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli, "Depth mapping using projected patterns," 2008. WO Patent WO/2008/120,217.

[63] D. Lanman and G. Taubin, "Build your own 3D scanner: 3D photograhy for beginners," in *SIGGRAPH '09: ACM SIGGRAPH 2009 courses*, (New York, NY, USA), pp. 1–87, ACM, 2009.

[64] M. Koch and M. Kaehler, "Combining 3d laser-scanning and close-range photogrammetry-an approach to exploit the strength of both methods," in *Making History Interactive. Computer Applications and Quantitative Methods in Archeology Conference*, pp. 22–26, 2009.

[65] P. G. Research, "Stereo accuracy and error modeling, http://www.ptgrey.com/," April 2004.

[66] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 519–528, accessed September 2006.

[67] C. Zach, M. Sormann, and K. Karner, "High-performance multi-view reconstruction," *3D Data Processing Visualization and Transmission, International Symposium on*, pp. 113–120, 2006.

[68] S. Gibson, P. Coe, A. Mitra, D. Howell, and R. Nickerson, "Coordinate measurement in 2-D and 3-D geometries using frequency scanning interferometry," *Optics and lasers in engineering*, vol. 43, no. 7, pp. 815–831, 2005.

[69] H. Liang, B. Peric, M. Hughes, A. Podoleanu, M. Spring, and S. Roehrs, "Optical coherence tomography in archaeological and conservation science–a new emerging field," in *Proc. of SPIE Vol*, vol. 7139, pp. 713915–1, 2008.

[70] L. Tian, N. Loomis, J. Domínguez-Caballero, and G. Barbastathis, "Quantitative measurement of size and three-dimensional position of fast-moving bubbles in air-water mixture flows using digital holography," *Applied optics*, vol. 49, no. 9, pp. 1549–1554, 2010.

[71] A. Pelagotti, M. Paturzo, A. Geltrude, M. Locatelli, R. Meucci, P. Poggi, and P. Ferraro, "Digital holography for 3D imaging and display in the IR range: challenges and opportunities," *3D Research*, vol. 1, no. 4, pp. 1–10, 2010.

[72] M. Hrebesh, Y. Watanabe, and M. Sato, "Profilometry with compact single-shot low-coherence time-domain interferometry," *Optics Communications*, vol. 281, no. 18, pp. 4566–4571, 2008.

[73] U. Kumar, B. Bhaduri, M. Kothiyal, and N. Mohan, "Two-wavelength micro-interferometry for 3-D surface profiling," *Optics and Lasers in Engineering*, vol. 47, no. 2, pp. 223–229, 2009.

[74] A. Podoleanu, "Optical coherence tomography," *British journal of radiology*, vol. 78, no. 935, pp. 976–988, 2005.

[75] A. Bradu, L. Neagu, and A. Podoleanu, "Extra long imaging range swept source optical coherence tomography using re-circulation loops," *Optics Express*, vol. 18, no. 24, pp. 25361–25370, 2010.

[76] J. Rosen, B. Katz, and G. Brooker, "Review of three-dimensional holographic imaging by Fresnel incoherent correlation holograms," *3D Research*, vol. 1, no. 1, pp. 28–35, 2010.

[77] R. Costa, R. Braga, B. Oliveira, E. Silva, T. Yanagi, and J. Lima, "Sensitivity of the moiré technique for measuring biological surfaces," *Biosystems Engineering*, vol. 100, no. 3, pp. 321–328, 2008.

[78] W. Ryu, Y. Kang, S. Baik, and S. Kang, "A study on the 3-D measurement by using digital projection moiré method," *Optik-International Journal for Light and Electron Optics*, vol. 119, no. 10, pp. 453–458, 2008.

[79] F. Mohammadi, K. Madanipour, and A. Rezaie, "Application of digital phase shift moiré to reconstruction of human face," in *UKSim Fourth European Modelling Symposium on Computer Modelling and Simulation*, pp. 306–309, IEEE, 2010.

[80] P. Tankam, Q. Song, J.-c. L. Mayssa Karrayand, J. M. Desse, and P. Picart, "Real-time three-sensitivity measurements based on three-color digital Fresnel holographic interferometry incoherent correlation holograms," *OPTICS LETTERS*, vol. 35, no. 12, 2010.

[81] G. Vogiatzis and C. Hernández, "Practical 3D reconstruction based on photometric stereo," *Computer Vision*, pp. 313–345, 2010.

[82] E. Prados and O. Faugeras, "Shape from shading," in *Handbook of Mathematical Models in Computer Vision* (Y. C. N. Paragios and O. Faugeras, eds.), ch. 23, pp. 375–388, Springer, 2006.

[83] G. Lippmann, "Épreuves réversibles donnant la sensation du relief," *Journal de Physique Théorique et Appliquée*, vol. 7, no. 1, pp. 821–825, 1908.

[84] F. Okano, H. Hoshino, J. Arai, and I. Yuyama, "Real-time pickup method for a three-dimensional image based on integral photography," *Appl. Opt.*, vol. 36, pp. 1598–1603, Mar 1997.

[85] B. Tavakoli, M. Daneshpanah, B. Javidi, and E. Watson, "Performance of 3D integral imaging with position uncertainty," *Optics Express*, vol. 15, no. 19, pp. 11889–11902, 2007.

[86] D. Chaikalis, N. Sgouros, and D. Maroulis, "A real-time FPGA architecture for 3D reconstruction from integral images," *Journal of Visual Communication and Image Representation*, vol. 21, no. 1, pp. 9–16, 2010.

[87] R. Sahay and A. Rajagopalan, "Dealing with parallax in shape-from-focus," *Image Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 558–569, 2011.

[88] R. Minhas, A. Mohammed, Q. Wu, and M. Sid-Ahmed, "3D shape from focus and depth map computation using steerable filters," *Image Analysis and Recognition*, pp. 573–583, 2009.

[89] M. Ahmad, "Focus measure operator using 3D gradient," in *ICMV 07*, pp. 18–22, IEEE, 2007.

[90] Y. An, G. Kang, I. Kim, H. Chung, and J. Park, "Shape from focus through laplacian using 3D window," in *2008 Second International Conference on Future Generation Communication and Networking*, pp. 46–50, IEEE, 2008.

[91] A. Thelen, S. Frey, S. Hirsch, and P. Hering, "Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation," *Image Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 151–157, 2009.

[92] P. Favaro, S. Soatto, M. Burger, and S. Osher, "Shape from defocus via diffusion," *IEEE transactions on pattern analysis and machine intelligence*, pp. 518–531, 2008.

[93] F. Galasso and J. Lasenby, "Shape from texture via fourier analysis," *Advances in Visual Computing*, pp. 803–814, 2008.

[94] J. Todd and L. Thaler, "The perception of 3D shape from texture based on directional width gradients," *Journal of vision*, vol. 10, no. 5, 2010.

[95] A. Lobay and D. Forsyth, "Shape from texture without boundaries," *International Journal of Computer Vision*, vol. 67, no. 1, pp. 71–91, 2006.

[96] A. Loh and R. Hartley, "Shape from non-homogeneous, non-stationary, anisotropic, perspective texture," in *Proc. of the BMVC*, pp. 69–78, Citeseer, 2005.

[97] Y. Sheikh, N. Haering, and M. Shah, "Shape from dynamic texture for planes," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, pp. 2285–2292, IEEE, 2006.

[98] G. S. Cheok, K. S. Saidi, M. Franaszek, J. J. Filliben, and N. Scott, "Characterization of the range performance of a 3D imaging system," Tech. Rep. NIST TN - 1695, NIST, 2011.

[99] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern recognition*, vol. 43, no. 8, pp. 2666–2680, 2010.

[100] M. Schaffer, M. Grosse, and R. Kowarschik, "High-speed pattern projection for three-dimensional shape measurement using laser speckles," *Applied optics*, vol. 49, no. 18, pp. 3622–3629, 2010.

[101] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (TOF) cameras: A survey," *Sensors Journal, IEEE*, no. 99, pp. 1–1, 2011.

[102] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt, "3D shape scanning with a time-of-flight camera," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1173–1180, 2010.

[103] R. Lange, P. Seitz, A. Biber, and R. Schwarte, "Time-of-flight range imaging with a custom solid-state image sensor," in *Proceedings of SPIE, the International Society for Optical Engineering*, vol. 3823, pp. 180–191, Society of Photo-Optical Instrumentation Engineers, 1999.

[104] B. Büttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger, "CCD/CMOS lock-in pixel for range imaging: Challenges, limitations and state-of-the-art," in *Proceedings of the 1st Range Imaging Research Day*, pp. 21–32, 2005.

[105] M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, "High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection," *Optics Letters*, vol. 36, no. 16, pp. 3097–3099, 2011.

[106] X. Su and Q. Zhang, "Dynamic 3-D shape measurement method: A review," *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 191–204, 2010.

[107] S. Nedevschi, R. Danescu, D. Frentiu, T. Marita, F. Oniga, C. Pocol, R. Schmidt, and T. Graf, "High accuracy stereo vision system for far distance obstacle detection," in *Intelligent Vehicles Symposium, 2004 IEEE*, pp. 292–297, IEEE, 2004.

[108] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P. Torr, "Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming," *International Journal of Computer Vision*, vol. 71, no. 1, pp. 89–110, 2007.

[109] T. Latychevskaia and H. Fink, "Solution to the twin image problem in holography," *Physical Review Letters*, vol. 98, no. 23, p. 233901, 2007.

[110] L. Bieman and K. Harding, "3D imaging using a unique refractive optic design to combine moiré and stereo," in *Proceedings of SPIE*, vol. 3204, p. 2, 1997.

[111] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Structured light 3d scanning in the presence of global illumination," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 713 –720, june 2011.

[112] J. Geng, "Structured-light 3D surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.

[113] I. Popov, S. Onuh, and K. Dotchev, "Dimensional error analysis in point cloud-based inspection using a non-contact method for data acquisition," *Measurement Science and Technology*, vol. 21, p. 075303, 2010.

[114] J. Valença, E. Júlio, and H. Araújo, "Applications of photogrammetry to structural assessment," *Experimental Techniques*, 2011.

[115] J. Dirckx, J. Buytaert, and S. Van der Jeught, "Implementation of phase-shifting moiré profilometry on a low-cost commercial data projector," *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 244–250, 2010.

[116] R. Lange, *3D time-of-flight Distance Measurement with Custom Solid-state Image Sensors in CMOS/CCD-Technology*. PhD thesis, University of Siegen, 2000.

[117] F. Brémand, J. Huntley, T. Widjanarko, and P. Ruiz, "Hyperspectral interferometry for single-shot absolute measurement of 3-D shape and displacement fields," *EPJ Web of Conferences*, vol. 6, 2010.

[118] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[119] P. Garg, N. Aggarwal, and S. Sofat, "Vision based hand gesture recognition," *World Academy of Science, Engineering and Technology*, vol. 49, pp. 972–977, 2009.

[120] G. Murthy and R. Jadon, "A review of vision based hand gestures recognition," *International Journal of Information Technology*, vol. 2, no. 2, pp. 405–410, 2009.

[121] M. Del Rose and C. Wagner, "Survey on classifying human actions through visual sensors," *Artificial Intelligence Review*, pp. 1–11, 2011.

[122] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 677–695, 1997.

[123] H. Stern, J. Wachs, and Y. Edan, "Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors," *Int. J of Semantic Computing. Special Issue on Gesture in Multimodal Systems*, 2008.

[124] J. LaViola, "A survey of hand posture and gesture recognition techniques and technology," *Brown University, Providence, RI*, 1999.

[125] M. Zobl, M. Geiger, K. Bengler, and M. Lang, "A usability study on hand gesture controlled operation of in-car devices," *Abridged Proceedings, HCI*, pp. 5–10, 2001.

[126] F. Althoff, R. Lindl, L. Walchshausl, and S. Hoch, "Robust multimodal hand-and head gesture recognition for controlling automotive infotainment systems," *VDI BERICHTE*, vol. 1919, p. 187, 2005.

[127] S. Sreedharan, E. S. Zurita, and B. Plimmer, "3D input for 3D worlds," in *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces*, OZCHI '07, (New York, NY, USA), pp. 227–230, ACM, 2007.

[128] D. Bowman, S. Coquillart, B. Froehlich, M. Hirose, Y. Kitamura, K. Kiyokawa, and W. Stuerzlinger, "3D user interfaces: New directions and perspectives," *Computer Graphics and Applications, IEEE*, vol. 28, no. 6, pp. 20–36, 2008.