Hindawi
Journal of Sensors
Volume 2017, Article ID 2072951, 24 pages
https://doi.org/10.1155/2017/2072951

Research Article

Multimodal Sensing Interface for Haptic Interaction

Carlos Diaz and Shahram Payandeh

Network Robotics and Sensing Laboratory, School of Engineering Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6

Correspondence should be addressed to Shahram Payandeh; payandeh@sfu.ca

Received 28 September 2016; Revised 13 December 2016; Accepted 12 January 2017; Published 4 April 2017

Academic Editor: Ivan A. Burenkov

This paper investigates the integration of a multimodal sensing system for exploring limits of vibrato tactile haptic feedback when interacting with 3D representation of real objects. In this study, the spatial locations of the objects are mapped to the work volume of the user using a Kinect sensor. The position of the user's hand is obtained using the marker-based visual processing. The depth information is used to build a vibrotactile map on a haptic glove enhanced with vibration motors. The users can perceive the location and dimension of remote objects by moving their hand inside a scanning region. A marker detection camera provides the location and orientation of the user's hand (glove) to map the corresponding tactile message. A preliminary study was conducted to explore how different users can perceive such haptic experiences. Factors such as total number of objects detected, object separation resolution, and dimension-based and shape-based discrimination were evaluated. The preliminary results showed that the localization and counting of objects can be attained with a high degree of success. The users were able to classify groups of objects of different dimensions based on the perceived haptic feedback.

1. Introduction

Several computer interfaces have been designed to improve the interaction between humans and computers. The traditional Graphical User Interface (GUI) combined with the mouse and keyboard provided a vast improvement in computers and allowed for tremendous growth in the computer industry. Yet, the new trends in computer development point toward ubiquitous portable devices that require different interface paradigms. Examples of Human-Computer Interfaces are touch screens for tablets and cellular phones, hand and body gestures for gaming platforms, and voice recognition for hand-free devices.

Multimodal sensing interface (MMSI) allows humans to interact with systems using several natural communication modes. These modes are referred to as the five human senses: sight, smell, touch, hearing, and taste. MMSIs move beyond the command-based interface enabling powerful, flexible, and more user-friendly interactive experiences.

This paper presents a design and implementation of a multimodal sensor interface for haptic interaction and exploration. The proposed system contains a 3D image acquisition module to obtain depth data of test objects on a scene. The MMSI includes the use of haptic feedback to provide remote perception of samples analyzed, a tracking and gesture recognition system for the hand of the user, and a graphical interface to display simplified 3D models of the observed scene.

The use of MMSI has several advantages according to some researchers. It has been found that when a stimulus excites several sensor modalities in synchrony, the information perception process improves dramatically and is less prone to errors, compared with the stimulation of individual channels [1, 2]. Another advantage is the possibility of alternating between single modes and multiple modes (multimode), offering the user alternatives. Properly designed MMSIs are easy to learn and use and can adapt to the requirements of users and situations. They offer the possibility of expanding the use of computers in a new spectrum of circumstances and users [1]. MMSIs are able to adapt to individual differences such as temporary or permanent handicaps [3].

The proposed design objective of the paper is to provide an experimental platform in the areas of visual and haptic interaction research. A hand tracking and hand gesture interface is used as the input modality. For the output modalities, the visual display and haptic display provide overlapping of data. These two modalities can be used independently or concurrently to accommodate individual user.

Remote sensing of objects could allow users to detect and avoid obstacles and enrich their perception. One of the potential applications of MMSIs is its usage as a navigation tool for visually impaired people. For example, Zöllner et al. [4] developed navigation MMI for blind users using in a depth camera attached to a helmet to measure the distance of objects in front of the user. The system provided haptic feedback using a wearable belt. Other authors have developed techniques to remotely locate objects and people using 3D image processing for partially or totally blinded people using visual, auditory, and tactile feedback [5–8].

Many MMSIs systems rely on hand gestures as an input. Several approaches have been proposed for hand gesture recognition. Lee et al. used stereovision for gesture detection [8], while Prisacariu and Reid [9] and Wang and Popović [10] developed color based segmentation techniques for bare hands and color gloves to deal with the background segmentation problem using video cameras.

In this paper we extend and present the design of a multimodal user interface based on a 3D sensor to gather spatial position information (Microsoft Kinect) from a scene, a haptic glove with vibrotactile feedback, and a hand locating system (Figure 1) [11]. The user can perceive the shape, location, and dimensions of the remote objects by moving the glove inside a scanning region. A marker detection camera-based module provides the location of the user hand (glove) for mapping the corresponding tactile feedback. Additionally a Gesture Recognition Subsystem was implemented providing the option to interact and control active elements such as computer interfaces or automated devices.

The rest of the paper is organized as follows. Section 2 shows a general description of the system with the associated sensing modality subsystems. Section 3 presents some discussions about the 3D user hand positioning. Section 4 presents hand posture measurement using retroreflective patterns; Section 5 presents the methodology for reconstructing and mapping the shape of the sensed depth image of the object in the reachable workspace of the user; Section 6 presents a design of haptic glove; Section 7 discusses the integration of the proposed sensing system; Section 8 presents a user study of our proposed system followed by discussion and conclusions remarks.

2. System Overview

The conceptual diagram of our proposed MMSIs system is shown in Figure 1. The system has three main components: (a) Bracelet Location Subsystem which is wore around the wrist of the user, (b) Depth Imaging Subsystem, and (c) haptic glove. A Kinect camera captures images of the sample objects placed within its field of view.

The diagram shown in Figure 1 can be further configured into four main components as shown in Figure 2. There are (a) *Bracelet Location Subsystem* (reflective bracelet + IR camera),

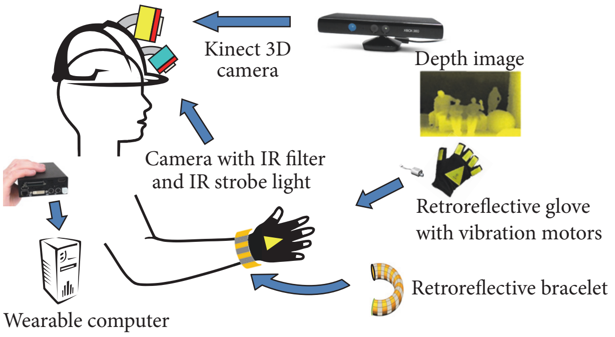

Kinect 3D camera
Depth image
Camera with IR filter and IR strobe light
Retroreflective glove with vibration motors
Retroreflective bracelet
Wearable computer

Figure 1: Overview of multimodal sensor interface for haptic interaction [12].

a video camera with embedded infrared illumination and the optical filters provide an image of the reflective bracelet; (b) *Gesture Subsystem* (globe + markers + IR camera), a gesture recognition module maps a given gesture made by the hand of the user to the graphical representation of the hand; and (c) *Depth Imaging Subsystem*, a depth sensor (Kinect camera) is used to provide the system with information about the environment. In the depth image, distances between camera and objects are encoded using different grey levels in a two-dimensional image. The brighter regions represent objects closer to the sensor. This information encodes a representation of obstacles in front of the user; (d) *haptic glove*: the haptic feedback system consists of a series of dc micromotors attached to the user's glove. The actuators generate eccentric load vibration signals for every finger in the hand and the palm. The vibrotactile sensation will be controlled with the wearable computer. Several messages with variation in intensity and rhythm can be used to encode feedback information from the 3 previous subsystems. Combining the environmental description (from the Depth Imaging Subsystem) with the location and orientation of the user hand (from the Gesture Recognition Subsystem), the haptic system can provide a "virtual tactile screen" representing obstacles present in the surrounding area.

In order to realize an experimental laboratory setup for the proposed system, the original concept shown in Figure 1 was modified. Sensors, cameras, and infrared lighting elements were attached to a fixed platform to enable controlled and repeatable measurements as shown in Figure 3. The setup consists of a platform to sample objects, with a surface of 60 cm × 120 cm. On the top of the sampling table there is a supporting fixture to support the cameras and electronic elements.

The sensor support holds a Kinect camera to capture images of the sample objects placed on the platform. In the same figure the "object sampling region" is shown with the green dashed lines. The Kinect depth sensor is inclined at an angle with respect to the horizon, covering the sampling region of the table within its field of view. Objects placed on top of this region can then be measured and classified according to basic geometrical shapes by the depth subsystem. On the same supporting base where the Kinect depth sensor is located, there is a greyscale camera from Point Grey. An

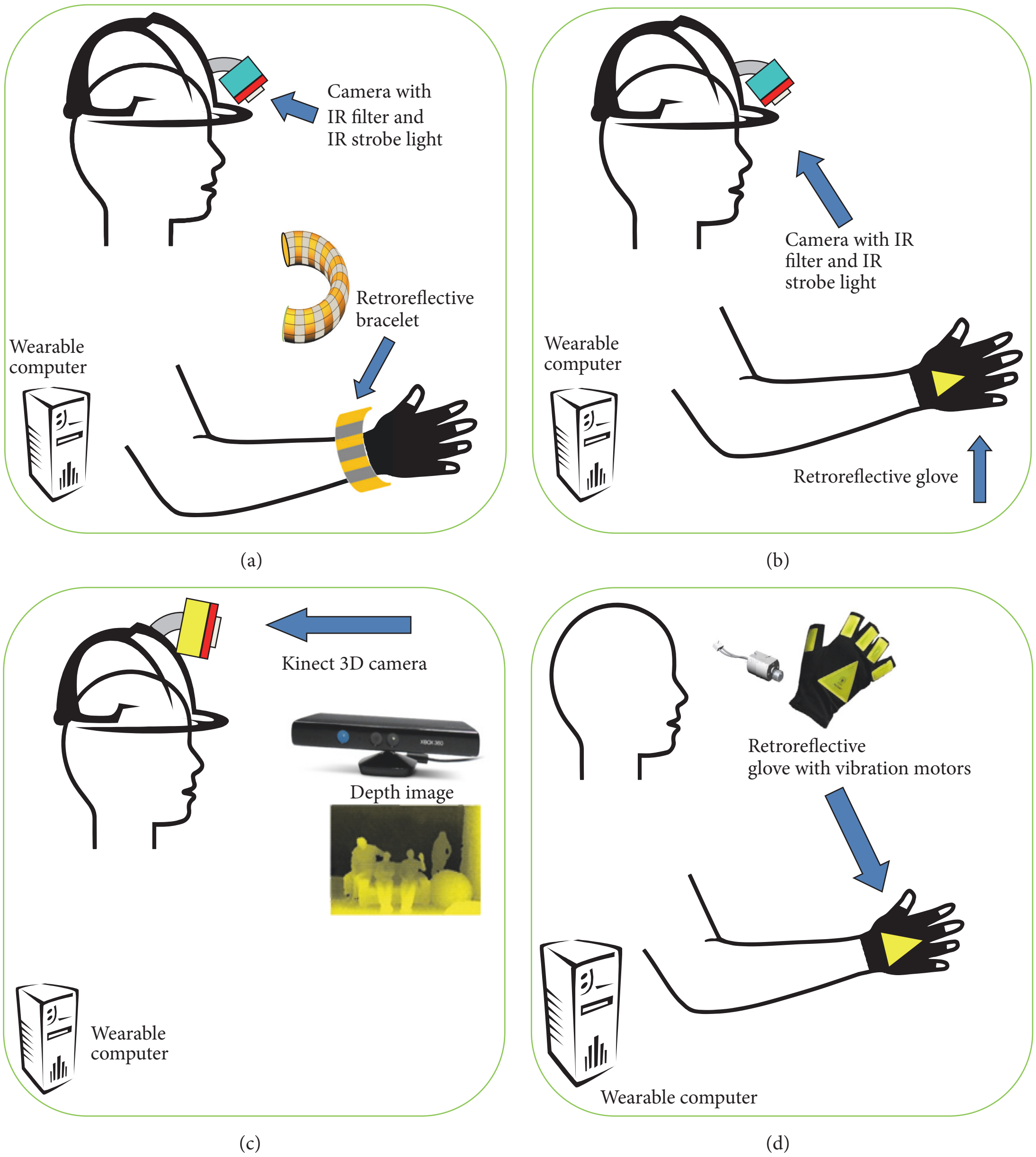


Figure 2: Multimodal user interface subsystems. (a) Hand bracelet locator; (b) gesture recognition; (c) depth sensing; and (d) haptic feedback.

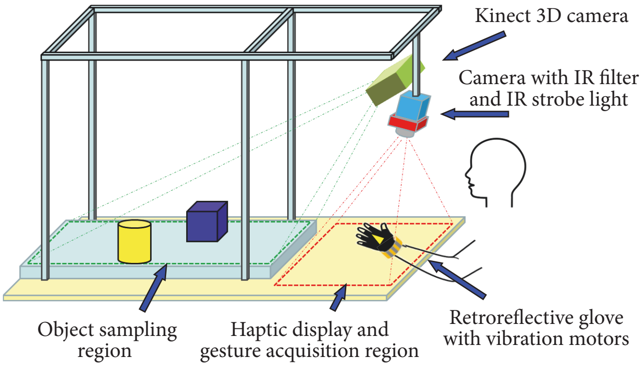
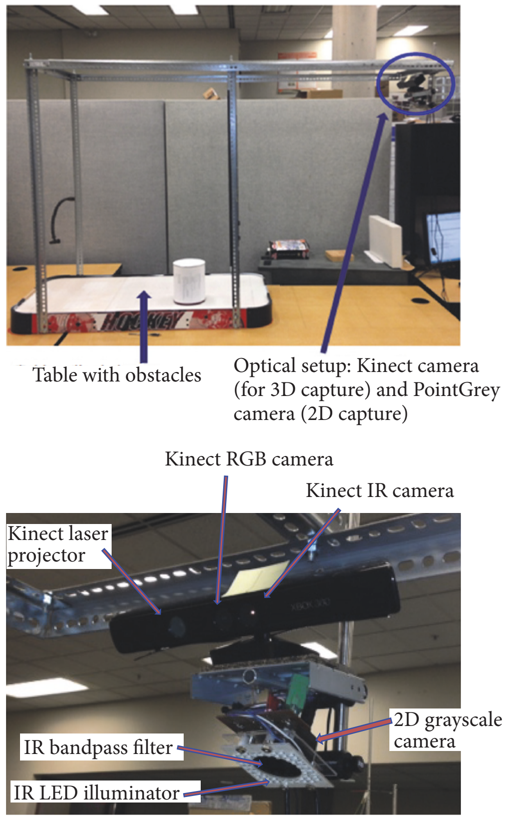


Figure 3: Schematic of the experimental setup.

infrared LED ring light is used to provide coaxial illumination for the camera (see Figure 4).

Attached below the Kinect sensor is a grayscale 2D camera with an infrared LED ring light which provides coaxial illumination. The field of view of this camera covers the “haptic display and gesture region” areas. Within this region, which is shown in the Figure 3 as red dashed lines, the position of the wrist and fingers of the user can be measured. This information can further be utilized as a part of a gesture recognition module. The combined measure for 3D reconstruction of the sensed object using Kinect and the bracelet position subsystem using camera and IR illumination is used in order to create haptic feedback.



Details of the optical setup

Figure 4: Images of the actual experimental setup [12].

Vibrotactile sensations are presented to the user as feedback data with the activation of the tactons on the glove. Each motor can be activated and modulated independently according to the required information to be displayed. By moving the hand in a scanning direction within the haptic region, it is possible to display haptic sensation corresponding to different areas of the sampling region. Data measured from objects placed on the sample platform can be encoded as tactile stimuli on the haptic display area.

3. Hand Bracelet Localization

The hand bracelet and a set of image processing tools are used to estimate the hand position in the 3D space. For the setup, a single channel greyscale camera with embedded lighting was used to obtain images of reflective markers on a bracelet. Figure 5 shows images of two reflective bracelets. The camera has a resolution of 752 × 480 pixels. A 4 mm lens with m12 mount was used with the camera. A compact LED lamp was used in combination with the camera. Lighting consists of a coaxial array of high power infrared LEDs in a ring-like distribution to control the illumination conditions. The model selected operates at a wavelength of 880 nm. This light source has the property of being invisible to the human eye, providing at the same time consistent illumination for the images to be captured with the camera. The imaging hardware provides digital controls for the physical properties of the camera. It is possible to adjust the gain of the camera to fixed (or programmable) values, avoiding the autogain function standard in most analog and digital cameras. Additionally, higher S/N ratios can be achieved by synchronizing the trigger of the camera and shutter time with the LED light. This setup allows a reduction in the amount of light coming from the environment and maximizing the amount of light from the ring light.

A band-pass infrared filter was used with the camera to improve the response of the system to the IR wavelengths by eliminating light sources out of the wavelength range. The use of coaxial lighting with the camera allows for a high reflection ratio from the reflective tape with respect to the environment. Most of the surrounding light will not be captured by the system, making it robust and stable for use in day or night conditions and indoors or outdoors.

A custom bracelet is used to obtain high contrast images to locate the bracelet in space. For this purpose, several bracelet designs covered with retroreflective tape were examined. Retroreflectors are surfaces that operate by returning light back to the light source along the same direction of light with a minimum scattering. A light ray is reflected back along a vector that is parallel yet opposite in direction to the light source.

Initial tests were performed with the imaging acquisition subsystem to determine the best parameters for the image acquisition. Figure 6(a) shows the image obtained using the ambient light present in our laboratory. From the image it is possible to obtain the glove (and markers). A considerable amount of computing resources to achieve adequate segmentation of the markers due to the cluttered background will be required. As mentioned before, a technique like background subtraction would allow for a good segmentation of the markers of the glove, but the required training and reinitialization would make this approach inconvenient in our case. Figure 6(b) shows the input image obtained with the use of the LED coaxial ring light. The use of coaxial illumination combined with retroreflectors increases the contrast of the reflector, simplifying the location of the markers on the image. Figure 6(c) shows the result of reducing the environmental light sources. This process is achieved by combining 2 effects. Originally, (in Figures 6(a) and 6(b)) the shutter time used in the camera was 16 ms. The shutter time refers to the time that the shutter remains open when taking a new image. Along with the aperture of the lens, it determines the amount of light that reaches the sensor. While Figure 6(a) shows the effect obtained just by using the environmental light, Figure 6(b) shows the effect of the environmental light in addition to the light coming from the controlled LED ring light. Considering that the LED light can be strobed for short periods of time (Ex 0.5 ms), we decided to limit the shutter time of the camera to be the same as the strobe light duration (0.5 ms) and in synchrony with it (both camera and lighting).

The second tool used to improve the contrast of the image was using an infrared filter on the camera. The Hoya R-72 infrared filter (which passes 720 nm and above) was used. The camera is sensitive to both white light and infrared light. By adding this IR band-pass filter to the optical setup, we could virtually eliminate all the remaining effects coming from

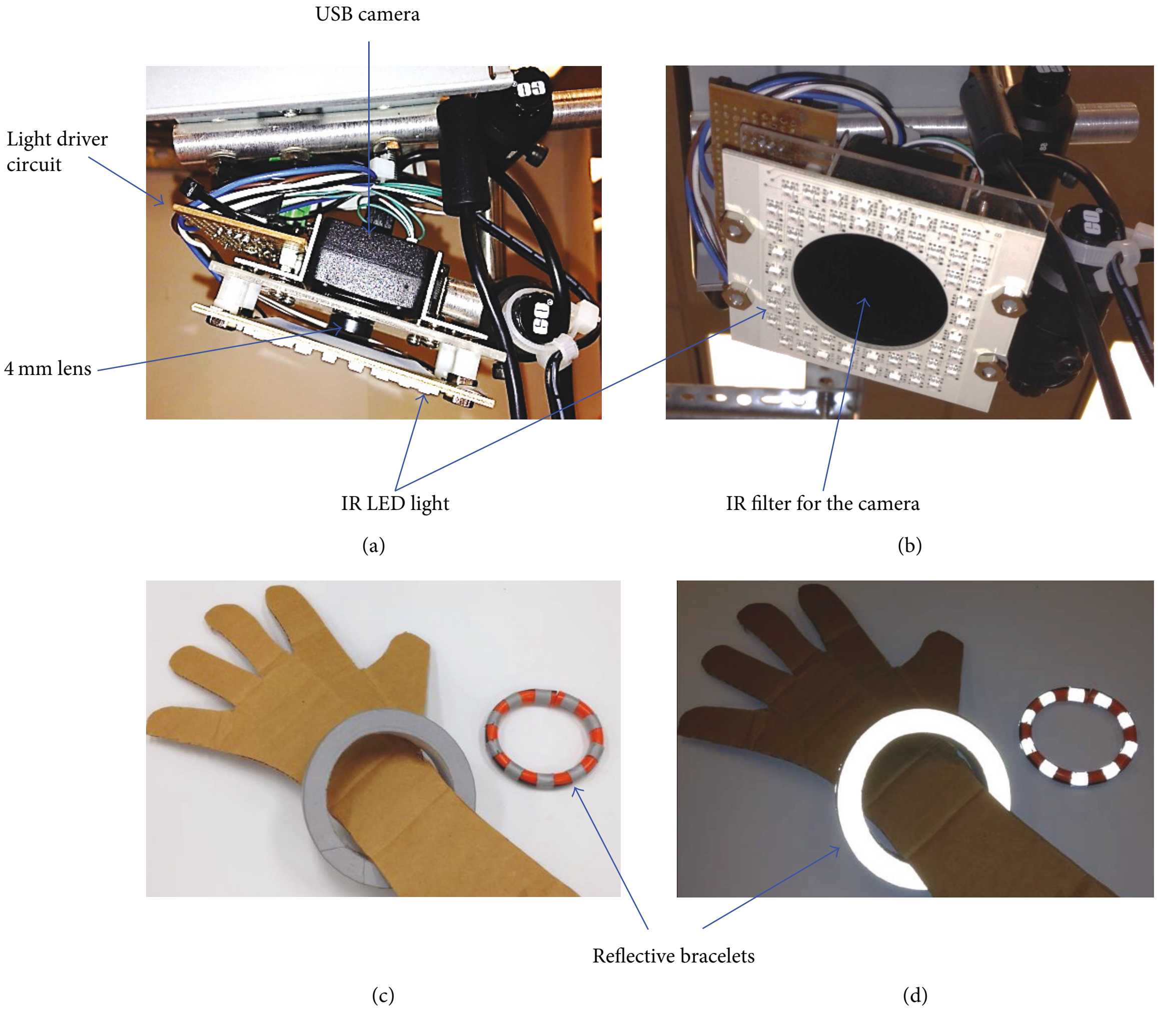
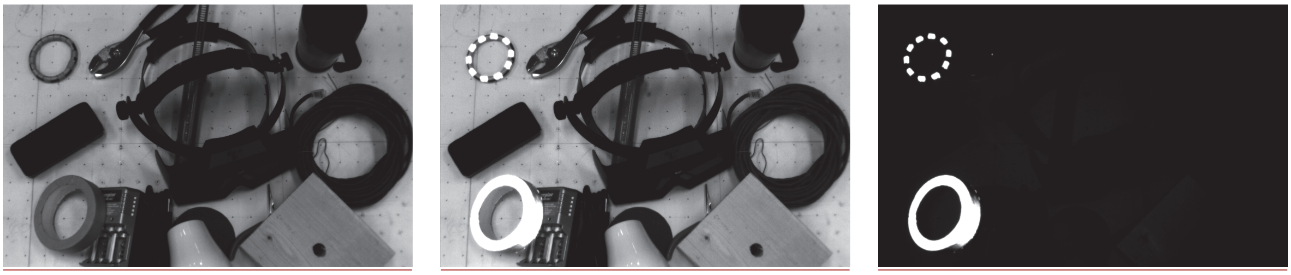


Figure 5: Image acquisition setup: USB camera, LED light, and IR filter from side view (a) and bottom view (b). Retroreflective markers are used on the bracelets. Picture without strobe light (c) and with strobe light [12].



Gain	296
Shutter time	16 ms
Strobe on (time)	0 ms
IR filter	No

(a)

Gain	296
Shutter time	16 ms
Strobe on (time)	0.5 ms
IR filter	No

(b)

Gain	167
Shutter time	0.5 ms
Strobe on (time)	0.5 ms
IR filter	Yes

(c)

Figure 6: Different light and camera setting for image acquisition. The markers can be observed, but the contrast between the desired objects (markers) and the background is low (a). Contrast improvement due to the strobe light (b). Contrast improvement due to the reduction of shutter time combined with IR filter on the camera lens.

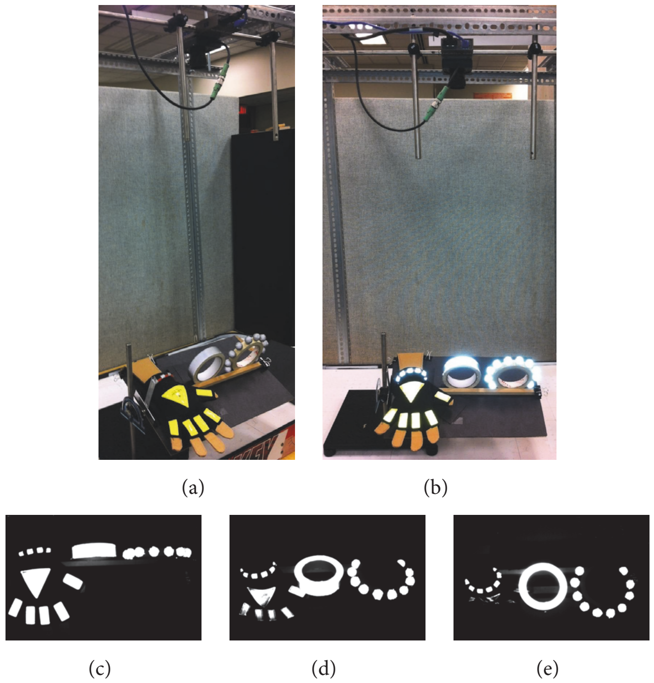

(a)
(b)
(c)
(d)
(e)

Figure 7: Different models for bracelet under test. Adjustable tilting platform to test response of reflective patterns under different incident angles (a) and (b). Image of 3 patterns at 0 degrees (c) (between the incident angle and the normal vector), image obtained at 40 degrees, and (d) image obtained at 80 degrees.

the surrounding light sources. The proposed optical setup is useful to obtain high contrast images with a controlled light source, additionally providing a very high S/N ratio (see Figure 6(c)).

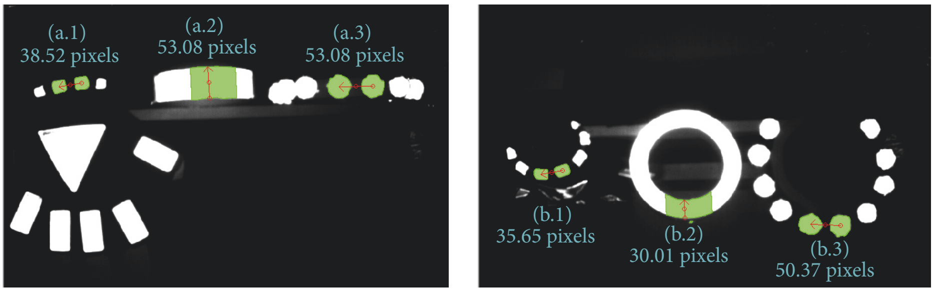
Retroreflectors, also called cataphotes, are devices or surfaces that reflect light back to its source with a minimum scattering of light. They are usually built using total internal reflection corner cubes (usually found in bicycle reflectors) or glass microspheres. Our designed gloves are covered with markers using 3M Scotchlite reflective material made with microscopic corner cubes. When used in combination with collinear/coaxial cameras and light sources they provide higher gains compared with the natural diffuse reflectivity of the materials.

3.1. Hand Bracelet Design. The patterns on the bracelet act as an invariant feature to provide information useful to compute the distance between the bracelet and the camera. With this invariant information, it is possible to calculate the position of the bracelet (and the wrist of the user) in a 3D space representation, working as a location accessory with the glove; the goal of using the bracelet is providing data about the location of the wrist of the user in any finger configuration of the hand and in any orientation in space. We tested 3 variations of the bracelet using 3M retroreflective tape in 3 different patterns.

Figure 7(b) shows from left to right a compact model design with 7 mm diameter plastic ring segments with small repeating patterns every 21 mm along the length of the bracelet. There is another model built with a tape roll displaying a continuous pattern of 2.5 cm width. The last model tested is built with a pattern of reflecting beads of 2 cm in diameter with a constant separation of 2 cm. The image of the bracelets was obtained at different angles of inclination of the platform under the optical IR camera and lighting while at the same time maintaining a "constant" distance between the camera and the samples in every measure. A simple image processing program was developed to measure some "candidate features" to be selected as invariant in the final bracelet design.

Figure 8 shows 2 screens of the test program developed to measure the distance between 2 features in pixels on each tested bracelet. Figure 8(a) corresponds to the initial position with 0 degrees between the incidence light direction and the vector normal to the support surface while Figure 8(b) is the image observed when the angle of the platform is 80 degrees. In this Figure, the feature labeled (a.1) represents the distance in pixels between the centroids of two consecutive marker segments (the centermost ones). Feature labeled (a.2) is the measure of the width of the second bracelet (in pixels). Feature 3 is the distance between the centroids of the 2 centermost marker beads. There is a difference between features pairs (a.1 b.1), (a.2 b.2), and (a.3 b.3), with the difference being more evident in the case (a.2 b.2).

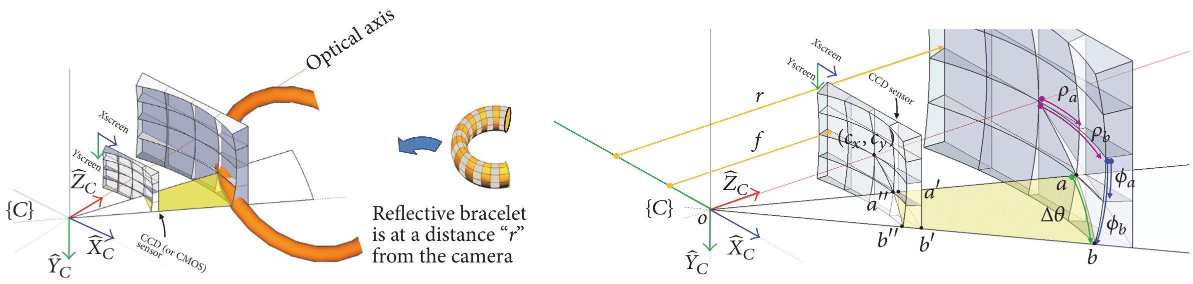
3.2. Hand Bracelet Localization. Given the pinhole camera model [13] and the known size of its radius, using spherical coordinates located at the camera frame, the objective is to compute the distance r with respect to the camera (Figure 9). The constant distance between consecutive markers of 21 mm



(a)

(b)

Figure 8: Measurements of "invariant features" in different models for the bracelet. Image at 0 degrees of inclination (a) and at 80 degrees of inclination (b).



(a)

(b)

Figure 9: Details considered for the calculation of distance r. The "constant" distance between 2 consecutive (and centermost) markers on the bracelet (points a and b) determines a unique angle $\Delta\theta$ for every distance r.

will be used to calculate the distance r between the bracelet and the camera. In Figure 9(b), points "a" and "b" represent the centroids of these two consecutive markers observed from the camera when the bracelet is at a distance r (both points are at a distance r from the camera).

The length of the circular arc with radius "r" between points "a" and "b" is $\widehat{ab} = r\Delta\theta$. Considering that the angle $\Delta\theta$ should be a small angle (angle of the invariant feature) compared with the field of view of the camera, it is useful to approximate the arc to a rectilinear segment using the simplification $r \approx B/\Delta\theta$. Additionally, these points can be projected on the surface of the CCD sensor as (a' and b'). Independently of the location of the bracelet in the field of view of the camera, it is possible to associate the radius r with the arc between points a and b.

From the geometry, the angles ρ and ϕ as a function of the projection coordinates on the CCD can be determined. At this point, an expression to compute angle $\Delta\theta$ as a function of ρ and ϕ is required. Given that the angles used are defined using a spherical coordinate system, we have utilized the spherical law of cosines to compute the angle $\Delta\theta$ resulting in the following expression:

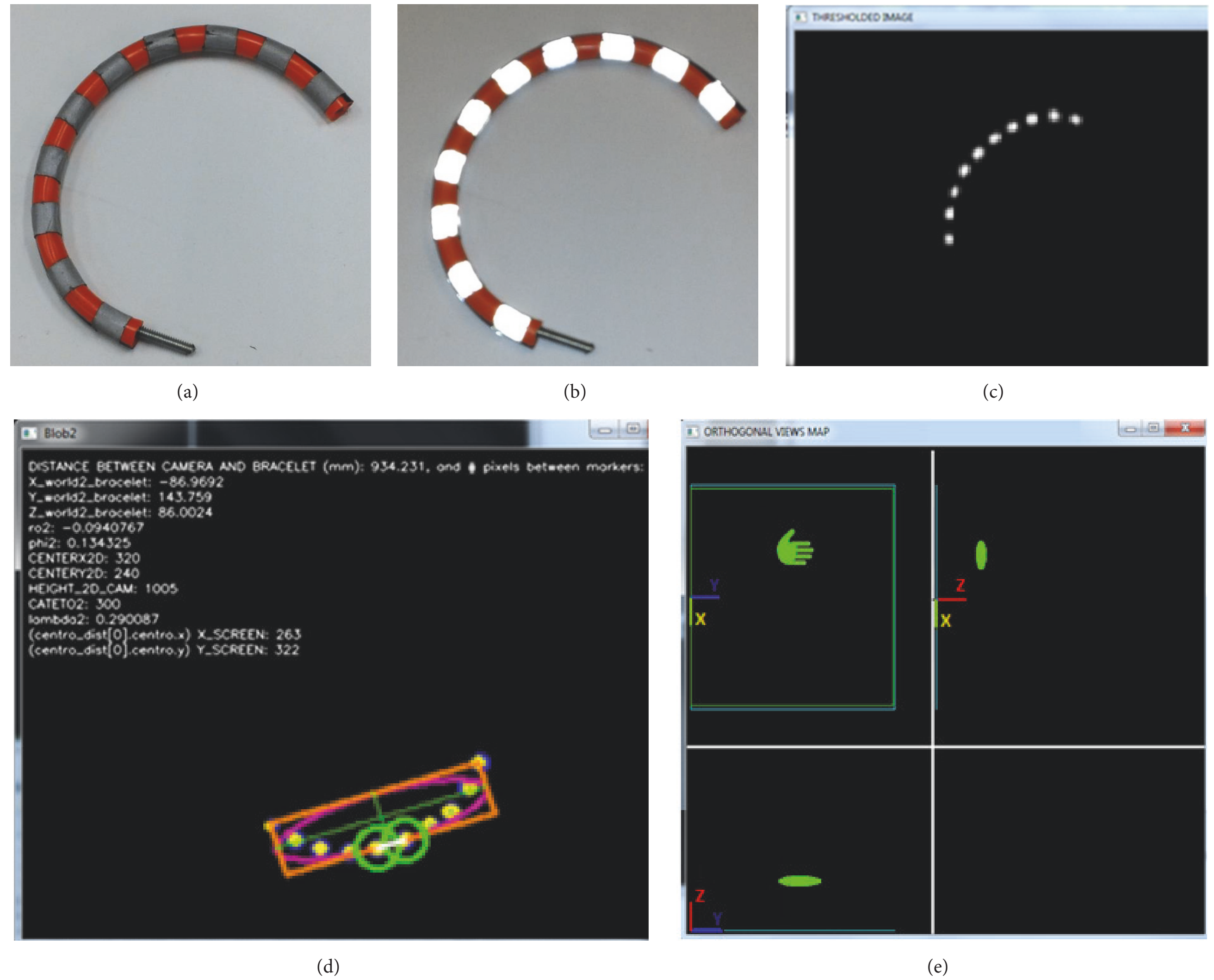
$$\Delta\theta = \cos^{-1}\left(\sin\phi_a \sin\phi_b + \cos\phi_a \cos\phi_b \cos\left(\rho_a - \rho_b\right)\right), \qquad (1)$$

where ϕ_a, ϕ_b, ρ_a, and ρ_b are computed from the spherical description [14]. For ϕ_b we use the coordinates of the project images b'_x and b'_y instead of a'_x and a'_y.

The location of the bracelet is used to determine the 3D position of the wrist of the user's hand. The following steps are applied in order to transform the original image of the user's hand wearing the bracelet to obtain the 3D location of the wrist (Figure 10).

Figure 10(a) shows the color image of the bracelet. Once the image is captured by the system using the mentioned optical parameters, the reflective sections respond with a very high contrast to the light stimuli as shown in Figure 10(c). Blob analysis is used to find individual segments of markers that belong to the bracelet. The sizes of the white regions are measured to verify they are in the range of minimum and maximum area values for these markers. Additionally, the index obtained dividing the major-axis-length by the minor-axis-length of every blob is used. This ratio allows for finding blobs with similar length and with (a ratio close to 1), which is a property of the bracelet segments.

In the next stage, bracelet segments are validated to see if they belong to the bracelet's vicinity. Every valid element belonging to the bracelet should be close to another bracelet element. With this evaluation, noisy areas, either too small or too isolated, can be eliminated. To find the required feature of interest from the remaining elements, the center



```
DISTANCE BETWEEN CAMERA AND BRACELET (mm): 934.231, and [illegible] pixels between markers:
X_world2_bracelet: -86.9692
Y_world2_bracelet: 143.759
Z_world2_bracelet: 86.0024
ro2: -0.0940767
phi2: 0.134325
CENTERX2D: 320
CENTERY2D: 240
HEIGHT_2D_CAM: 1005
CATETO2: 300
lambda2: 0.290087
(centro_dist[0].centro.x) X_SCREEN: 263
(centro_dist[0].centro.y) Y_SCREEN: 322
```

(a) (b) (c)

(d) (e)

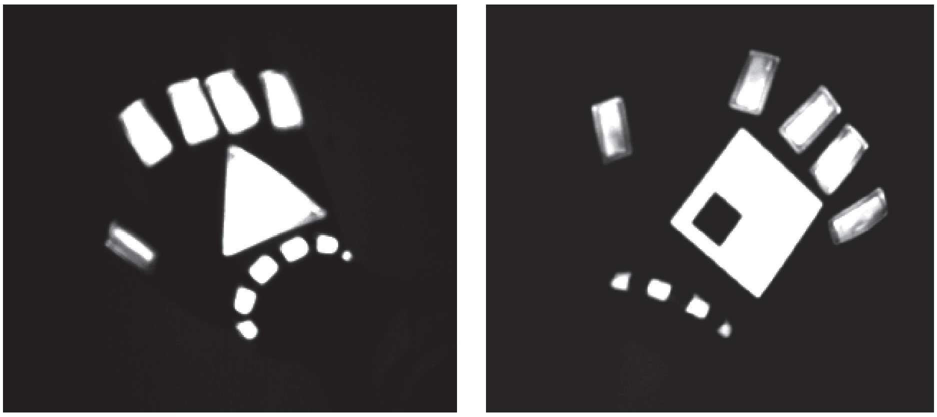
Figure 10: Screens of the image processing module to locate the bracelet. Bracelet under ambient light (a) and strobe light (b). Binary image showing the elements corresponding to the bracelet (c). Computation of the distance between bracelet and camera and the 3D location with respect to the camera frame $\{C\}$ and world frame $\{W\}$ (d). Using an orthogonal view with parallel projection, the estimated position of the wrist of the user hand is computed (e).

of the containing region of interest is determined. The two elements closer to the center of the region containing the bracelet elements should correspond with the two centermost markers on the bracelet and they should be the elements closest to the camera (as in Figure 9). These two elements are shown in Figure 10(d) enclosed in two small green circles. In Figure 10(e) an icon (small green hand) is placed to indicate the position of the wrist (and the inferred position of the hand) in a 3D volume. For this representation, parallel projection is used and the hand location is presented in planes X-Y, Y-Z, and X-Z.

4. Hand Posture Measurements

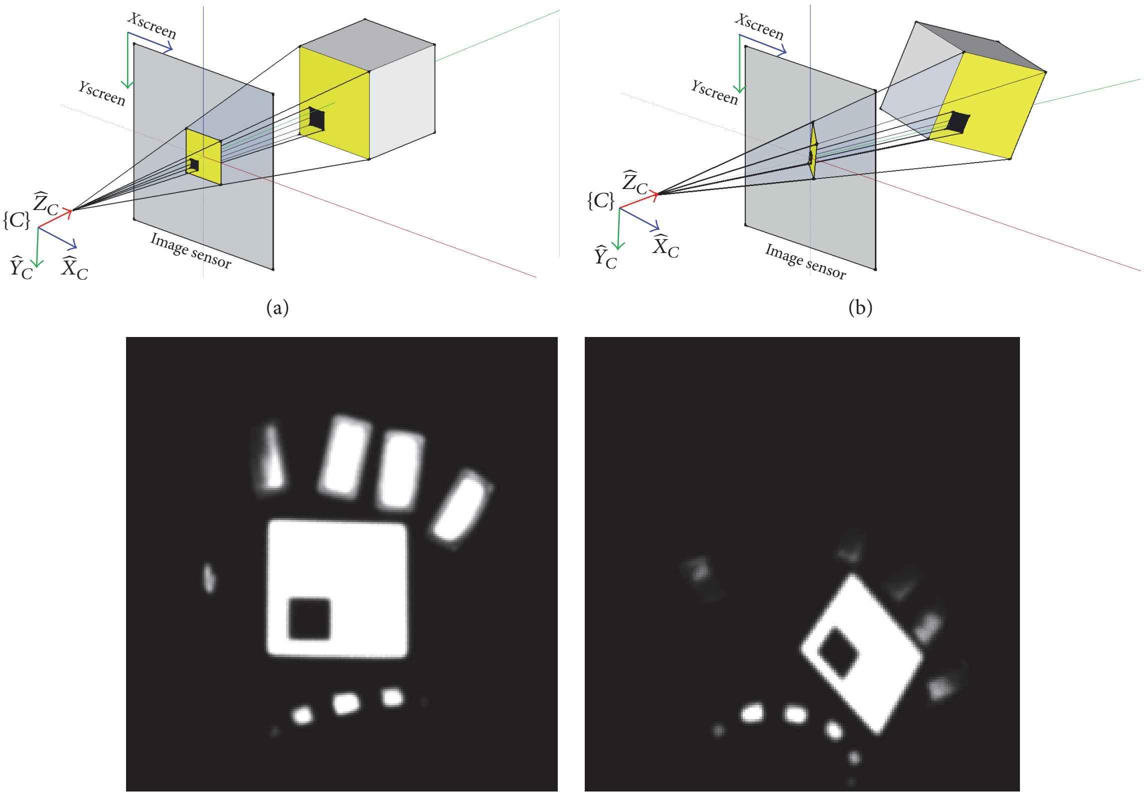
In combination with the use of the hand bracelet presented in the previous section, a glove enhanced with markers (see Figure 11) was designed to provide the system with information about the hand position and orientation and configuration of the fingers [12, 15–17]. Using a frame associated with the glove, the proposed design is able to estimate the position of the hand frame (H) with respect to the world frame (W). There exist other approaches for recognizing the posture of the hand of the user. For example, [18] has proposed an approach using the RFID technology which offers a cheap and unintrusive passive tags which can be easily attached to or interweaved into user clothes, which can be read by RFID antennas. [19] proposed a color patched wearable glove in which a color camera can be used to classify various basic postures of the hand. [20] proposed an approach for posture recognition of hands using a network of calibrated cameras.

In our study, shapes of the marker pattern which are used to describe the configuration of the hand are shown in Figure 11. The figure shows the image of the hand bracelets and a number of square, rectangular, and triangular markers.



(a) (b)

Figure 11: Two marker patterns on the gesture glove [12].



(a) (b)

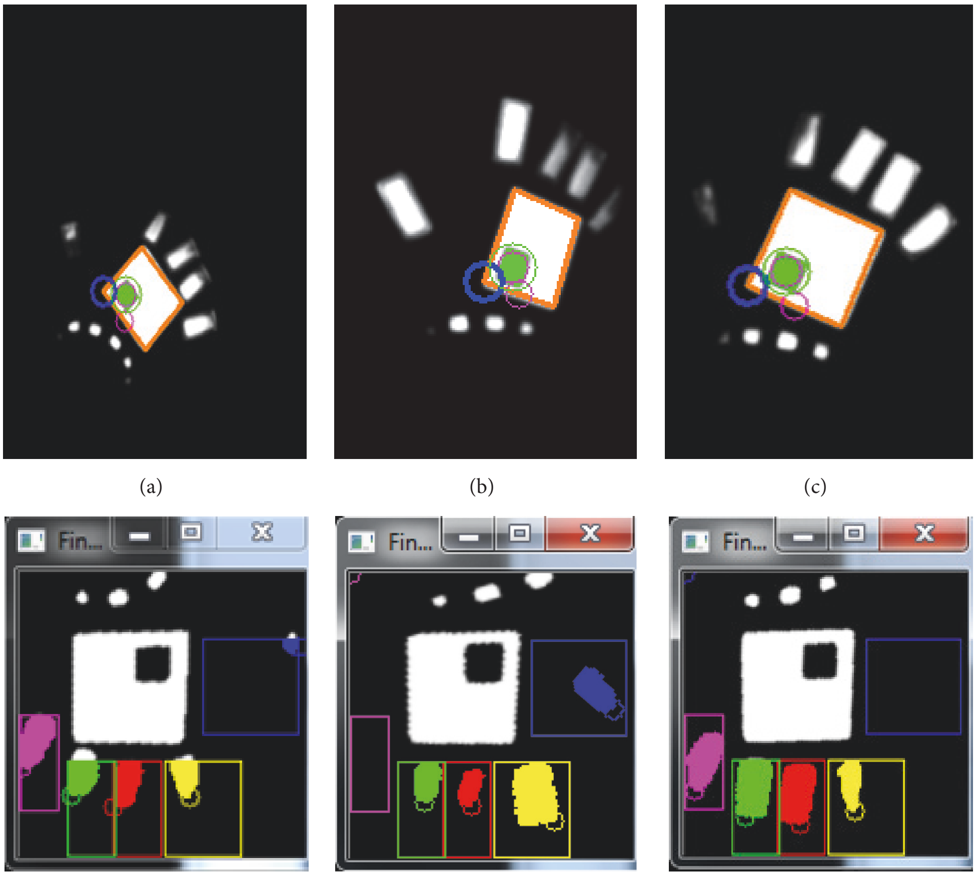
(c) (d)

Figure 12: Effect of perspective transformation in the project image. A square image (a) can be mapped to a trapezoid type image [12].

In this paper, the second pattern was used where a known square pattern was attached firmly on the back of the manipulating glove. Figure 12 shows the image of this marker under two different orientations of the hand.

The main objective of using a known pattern on the glove is to provide a normalized template in order to simplify image processing in locating the fingers [21]. Once the region corresponding to the square is located on the image, a novel approach which uses the four corners for defining the trapezoidal shape can be easily obtained. In this paper, these four noncollinear points are the minimum requirement needed in order to compute the transformation matrix H defined through projected images and homography [22]. The x, y coordinates of the points of the trapezoid are related to the four corners points of a normalized square by the matrix H with the equation $P_{dst} = HP_{src}$.

Homography matrix H has 8 degrees of freedom, meaning that in their computation only 8 of the 9 values are



(a)

(b)

(c)

(d)

(e)

(f)

Figure 13: The original image of the hand is analyzed to locate the trapezoid (corresponding to the hand) (a), (b), and (c). The smaller black square inside provides a reference to identify the corners. The corner points are used to estimate matrix H. Perspective "normalization" is obtained after applying H inverse, (d), (e), and (f).

required to fill the matrix H. With the 4 points of the corners of the trapezoid the set of equations to compute matrix H is complete. The inverse of the H matrix can be used to map the rotated spare marker back to its origin unrotated configuration. In this way, it is convenient to locate the fingers in an invariant configuration. Figure 13 shows this effect of perspective *rectification*.

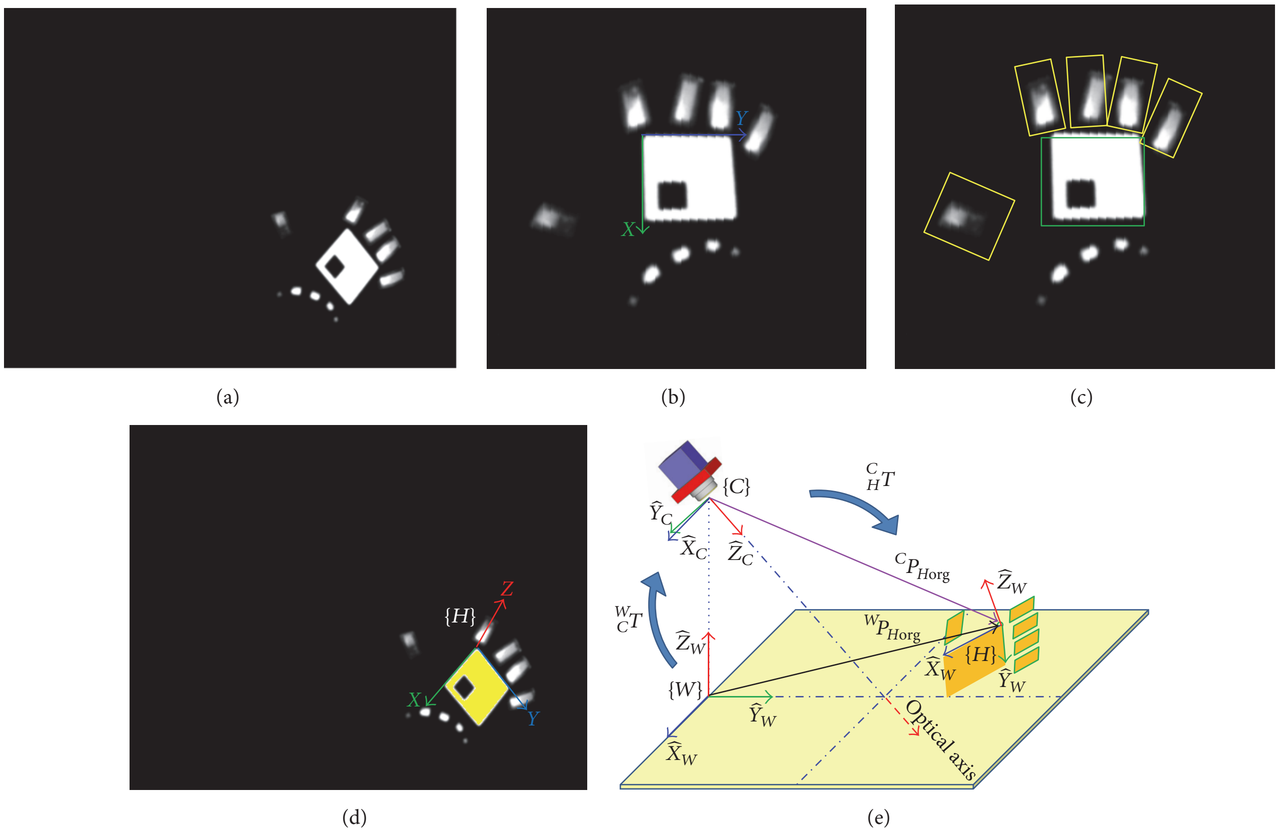
4.1. Positon and Orientation of the Hand. The Perspective Projection is a transformation from a tridimensional space to 2D. It can be represented by a 3×4 matrix P (also called camera matrix) such that $x = PX$, where X is a 4×1 vector of homogeneous coordinates with the coordinates of a point in the world and x is a 3×1 vector of homogeneous coordinates with point coordinates on the image plane. Camera matrix P can be decomposed in an extrinsic parameter matrix and extrinsic parameters.

P matrix is defined as $P = K[R \mid t]$, where K is the 3×3 intrinsic parameters matrix of the camera and $[R \mid t]$ is the 3×4 matrix of extrinsic parameters. Intrinsic parameters of the camera are using the method of [23], to result in f_x and f_y which are the focal distances in x and y direction. c_x and c_y are the principal point coordinates.

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (2)$$

The extrinsic parameters contain information about the position and orientation of a frame attached to the points in the world coordinates (Figure 14). Assuming that images obtained from a camera are subject to a homography transformation, Zhang [23] proposed an algorithm to compute the intrinsic and extrinsic parameters considering that the homography H can be expressed as $H = K[R \mid t]$.

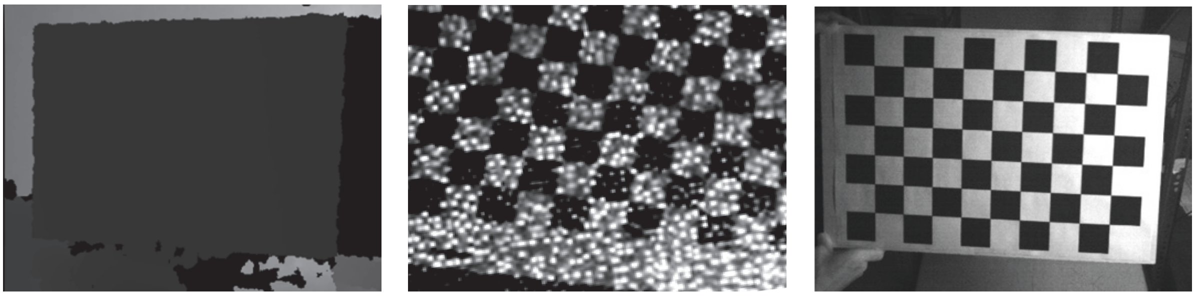
The extrinsic parameter matrix can be written as the description of the frame attached to the hand $\{H\}$ with respect to the camera frame $\{C\}$, by using homogeneous coordinates ${}_{H}^{C}T$. The relative rotation matrix of the hand frame with respect to the world frame can be easily defined.



(a) (b) (c)

(d) (e)

FIGURE 14: Extrinsic parameters represent the description of a frame attached to the hand $\{H\}$. The corners of the hand trapezoid are used to "rectify" the effect of perspective (b). Each finger has a location on the normalized hand (c). A frame is assigned to the trapezoid (d) and extrinsic parameters are computed. Extrinsic parameters define the transformation $^{C}_{H}T$ in (e).



(a) (b) (c)

FIGURE 15: Kinect images. Depth image of the chessboard used for calibration (a). Raw infrared image captured by the CMOS sensor of the Kinect (b) (illuminated with the laser projector.) In (c), the laser projector has been blocked and the chessboard was illuminated using a halogen lamp resulting in more even distribution of light.

5. Depth Image Processing

The Depth Imaging Subsystem utilizes a depth sensor (Kinect-I sensor), for obtaining spatial information in a form of (x, y, and z) coordinates of a pixel representation of the scene with respect to its reference coordinate frame. In order to calibrate the sensor, a 9 × 6 chessboard was used under IR imaging (Figure 15) [24].

A total of 25 images of the chessboard were used as the input for the calibration algorithm. The calibration algorithm results provide the focal length corresponding to the X and Y axes (f_x, f_y), the coordinates of the principal point, and the distortion coefficients. This information can be used to rectify images obtained from the depth sensor to compensate for the effects radial and tangential distortion of the lens. For example, for the sensor used in our study, we have obtained focal length (f_x, f_y) of 581.69 and 589.72, respectively.

5.1. Locating the Reference Frame of Kinect Sensor. Depth data obtained with the Kinect is used to locate objects in the Kinect

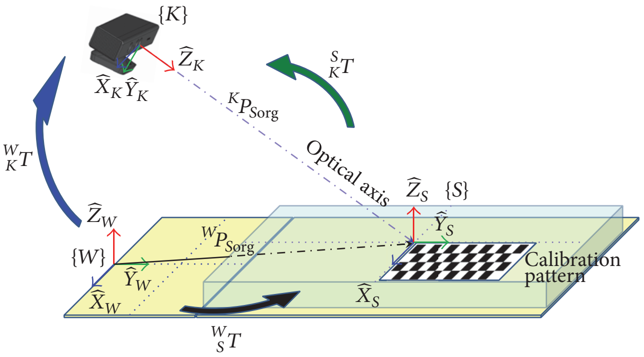


Figure 16: Procedure to find the position and orientation of the Kinect with respect to the world frame (${}^{W}_{K}T$). If frames ${}^{W}_{S}T$ and ${}^{S}_{K}T$ are known, then the transformation describing Kinect frame $\{K\}$ with respect to world frame $\{W\}$ is given by ${}^{W}_{K}T = {}^{W}_{S}T\,{}^{S}_{K}T$. Frame $\{S\}$ can be found by computing the extrinsic parameters with the calibration board in the center of the test table.

reference frame $\{K\}$. To find the description of these objects in a different frame, the transformation relating Kinect frame with respect to this reference frame is required. In this case, the same world frame $\{W\}$ was used to describe the samples. The definition of the world frame is arbitrary in this case and can be defined, for example, as a frame attached to the wearable head-gear shown in Figure 1. For the case study of this paper, we have set the world frame to be a location on the supporting table (see Figure 16). The Kinect frame description with respect to the world frame is represented by the transformation ${}^{W}_{K}T$. This mapping can be obtained through the following relationship: ${}^{W}_{K}T = {}^{S}_{K}T\,{}^{W}_{S}T$ (Figure 16).

The calibration algorithm was used to compute the extrinsic parameters of frame $\{S\}$ which are the description parameters of the frame attached to the chessboard $\{S\}$ with respect to the Kinect frame $\{K\}$. Figure 17(a) shows the example image used for extrinsic parameters calculations. Figure 17(b) shows the location of the X and Y axes of the frame $\{S\}$. The extrinsic parameters describe transformation ${}^{K}_{S}T$, defining frame $\{S\}$ with respect to Kinect frame $\{K\}$. The extrinsic parameters for the experimental results of this study are computed as

$$
{}^{K}_{S}T = \begin{bmatrix} 0.99 & -0.025 & 0.002 & -38.792 \\ -0.014 & -0.643 & -0.765 & -15.373 \\ 0.020 & 0.765 & -0.643 & 1701.31 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{K}_{S}R & {}^{K}P_{\text{Sorg}} \\ 0\ \ 0\ \ 0 & 1 \end{bmatrix}, \tag{3}
$$

where ${}^{K}P_{\text{Sorg}}$ is the location of the origin of frame $\{S\}$ with respect to $\{K\}$ and ${}^{K}_{S}R$ is the rotation matrix. To find ${}^{S}_{K}T$ we apply the following relationship:

$$
{}^{S}_{K}T = \begin{bmatrix} {}^{K}_{S}R^{T} & -{}^{K}_{S}R^{T}\,{}^{K}P_{\text{Sorg}} \\ 0\ \ 0\ \ 0 & 1 \end{bmatrix}, \tag{4}
$$

resulting in

$$
{}^{S}_{K}T = \begin{bmatrix} 0.99 & -0.014 & 0.020 & 2.928 \\ -0.025 & -0.643 & 0.765 & -1312.49 \\ 0.0025 & -0.765 & -0.643 & 1083.31 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{S}_{K}R & {}^{S}P_{\text{Korg}} \\ 0\ \ 0\ \ 0 & 1 \end{bmatrix}. \tag{5}
$$

In our setup, Vector ${}^{W}P_{\text{Sorg}}$ describes the position of the origin of frame $\{S\}$ with respect to frame $\{W\}$. Here, the magnitude of the X coordinate of the ${}^{W}P_{\text{Sorg}}$ vector is 0, the Y coordinate = 1375 mm, and the Z coordinate = 75 mm. Or,

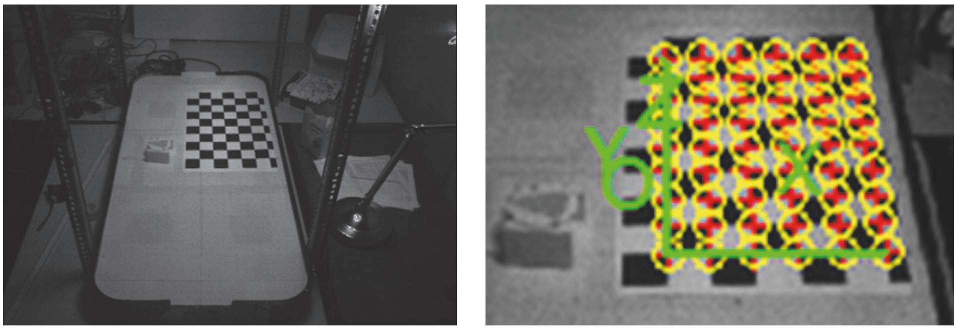
$$
{}^{W}_{S}T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1375.00 \\ 0 & 0 & 1 & 75.00 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{W}_{S}R & {}^{W}P_{\text{Sorg}} \\ 0\ \ 0\ \ 0 & 1 \end{bmatrix}, \tag{6}
$$

where we can now compute the matrix ${}^{W}_{K}T = {}^{S}_{K}T\,{}^{W}_{S}T$ as

$$
{}^{W}_{K}T = \begin{bmatrix} 0.99 & -0.014 & 0.020 & 2.928 \\ -0.025 & -0.64 & 0.765 & 62.508 \\ 0.002 & -0.76 & -0.643 & 1158.31 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{W}_{K}R & {}^{W}P_{\text{Korg}} \\ 0\ \ 0\ \ 0 & 1 \end{bmatrix}. \tag{7}
$$

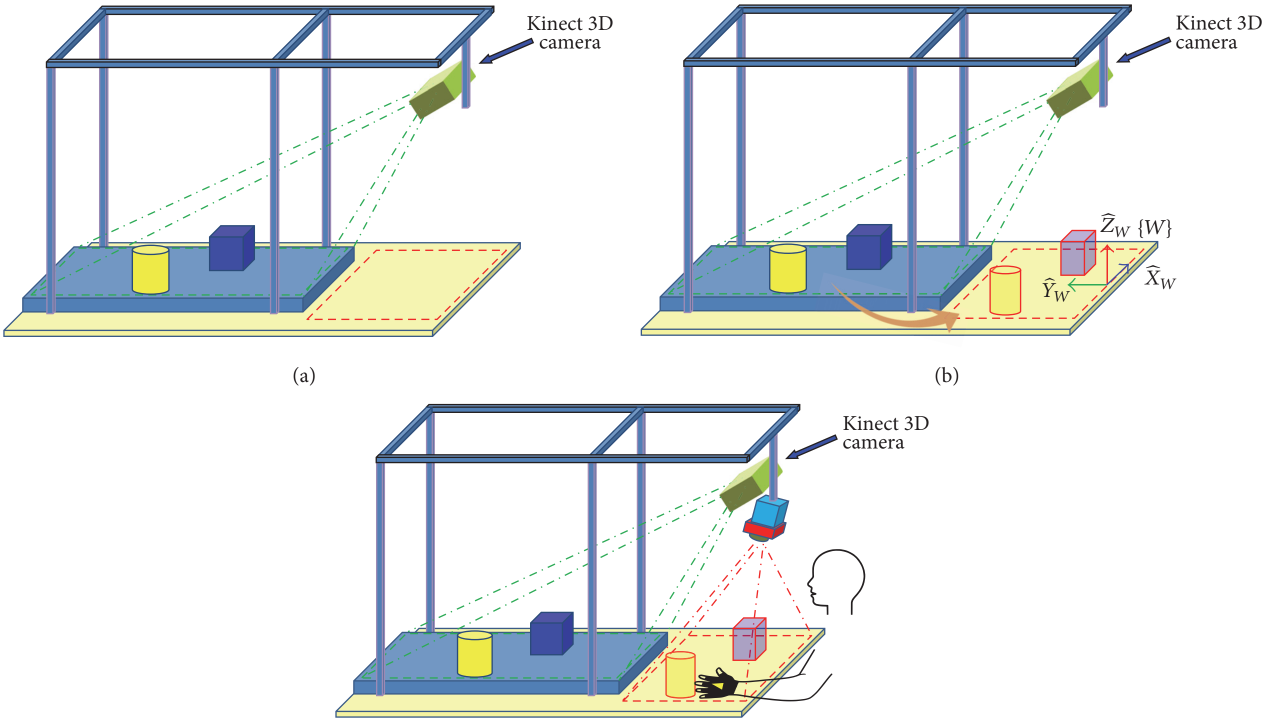
5.2. Object Localization. Here the objective is to identify and isolate the regions of interest inside the depth image which contains the objects. Once new samples of objects are placed on the test table, a basic background subtraction operation is carried out which allows for the elimination of the environmental cluttering resulting in a depth image of the objects of interest (Figure 18). Background subtraction has been shown to be an effective approach for segmenting objects in depth images [25]. It can also be used in a more general application setup of the proposed system when integrated with the wearable head-gear configuration of Figure 1 for segmenting people as they move in front of the depth sensor [26].

The grayscale images obtained from the depth sensor are encoded to represent distances between objects and the camera plane (Figure 19((a) and (b))). Due to the geometry of the 3D sensor and the reflective properties of some of the surfaces (i.e., matte or glossy surfaces), there are regions in the image that cannot be measured properly. In the Kinect-I image, the color white (RGB = 255, 255, and 255) is reserved for pixels that cannot be quantified in depth. Binary masks (shown in Figure 19((f) and (g))) are used to eliminate undetermined values in the image such as the edges of the testing table and the supporting structure using a binary



(a) (b)

Figure 17: Extrinsic parameters of a frame. The chessboard was placed in the center of the test table (a) as observed from the Kinect IR camera. With the use of the calibration routine it was possible to compute the extrinsic parameters describing the position and orientation of the frame attached to the calibration board. This frame (in the center of the test table) was named {S} (b).



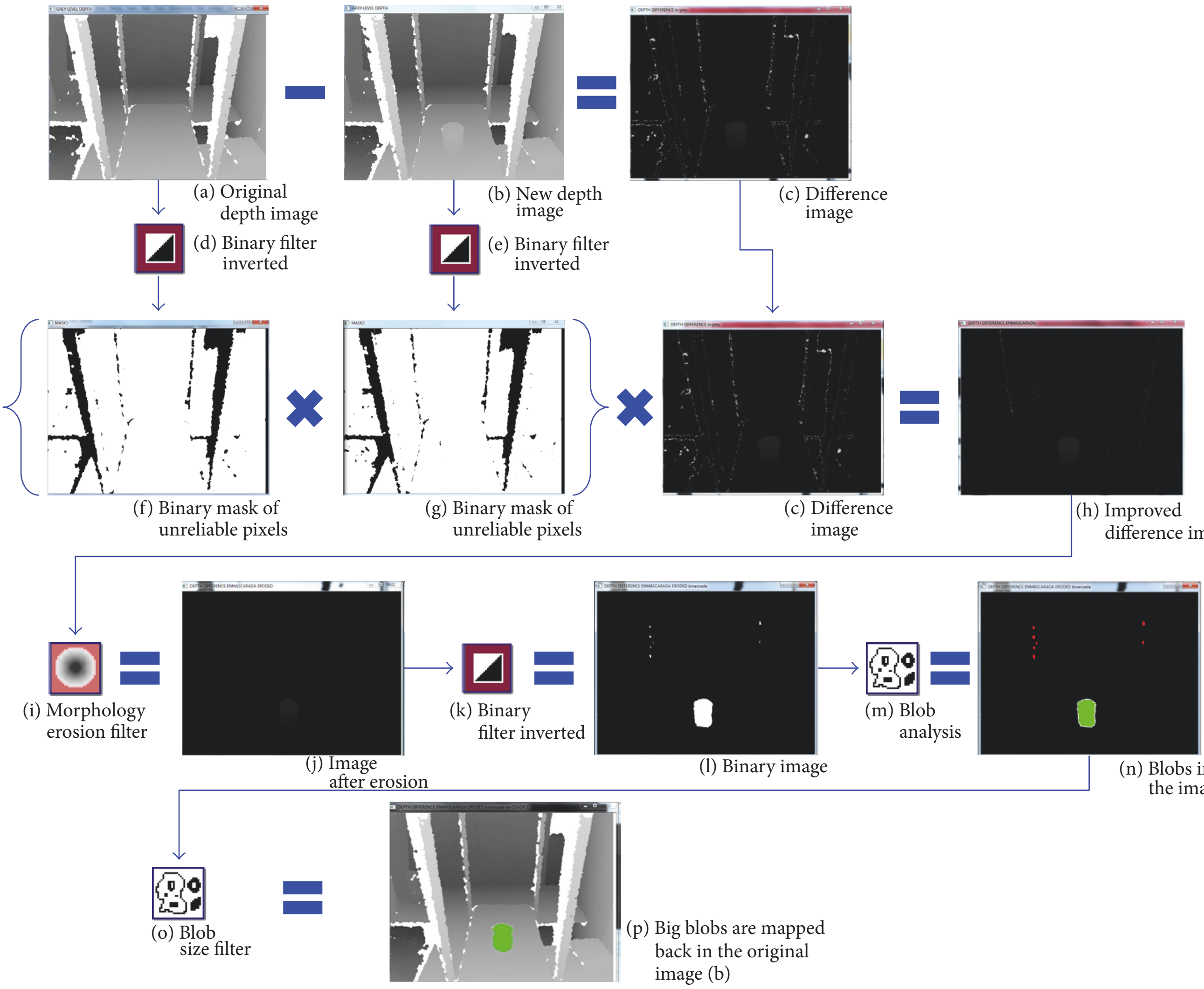
(a) (b) (c)

Figure 18: 3D depth image reconstruction procedure. (a) Objects placed on the test platform are located with the Kinect sensor. Dimensions of the objects are computed to describe them using parallelepiped or cylindrical volumes. In (b), a scaled representation of the objects is used to build virtual volumes in the workspace of the user. The virtual models are used as an input of the haptic module to provide tactile sensations when the virtual objects are penetrated by the hand of the user (c).

filter set to transform white pixels into black pixels. Masks in Figure 19((f) and (g)) are combined and multiplied to the difference image (Figure 19(c)) to eliminate noise edges of the test table and structure.

Several stages of morphology operators (erosion, dilation) are used to reduce the effect of noise in the regions of interest [27]. Blob analysis is used to obtain enhanced images of the regions of interest by filtering out regions that have very small areas (considered noise in the depth image) [28]. Finally, a blob filter allows for the selection of connected areas with significant areas whereas very small regions are eliminated. The resulting blobs are mapped onto the original depth image (green region in Figure 19(p)).

5.3. Object Description. In this stage of our case study, measurements of height and a set of points describing the basic shapes of objects are obtained [29]. The height of each test object is considered one of the main descriptors. The point (of the object) closer to the depth sensor is used to define the height of the entire object. Figure 20 shows an overview of the



(a) Original depth image

(b) New depth image

(c) Difference image

(d) Binary filter inverted

(e) Binary filter inverted

(f) Binary mask of unreliable pixels

(g) Binary mask of unreliable pixels

(c) Difference image

(h) Improved difference image

(i) Morphology erosion filter

(j) Image after erosion

(k) Binary filter inverted

(l) Binary image

(m) Blob analysis

(n) Blobs in the image

(o) Blob size filter

(p) Big blobs are mapped back in the original image (b)

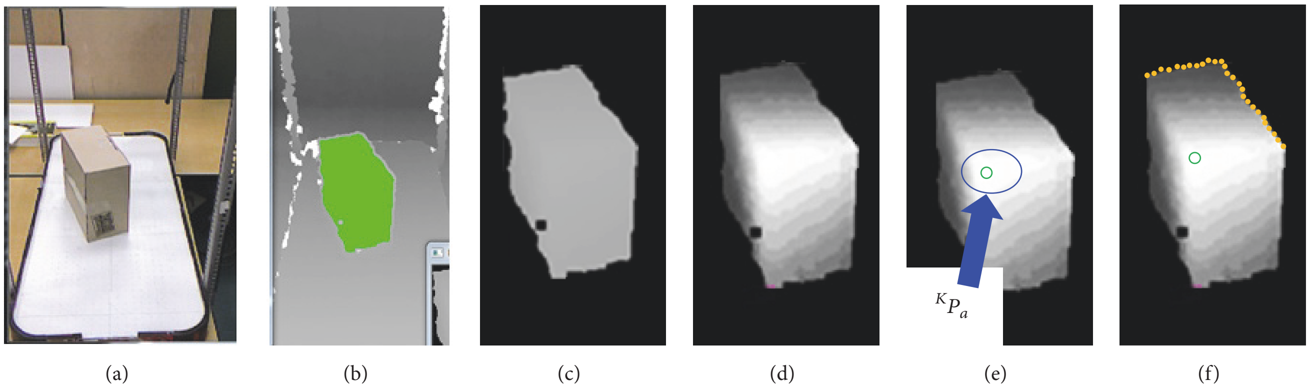
Figure 19: Object localization algorithm. Detailed steps in the object segmentation process. From a background image (a) and image of objects placed on the test table (b), it is possible to isolate the objects in the scene (p). Segmented objects are mapped back on the depth image using connected components (green mask on (p)).

object description algorithm. The objective is to determine the position of the corner (or edge) of the object closest to the depth sensor. Figure 20(a) contains a mask of the depth data corresponding to example object (the box). Histogram equalization remaps (Figure 20(d)) an image in the entire range (0–255), in order to increase the contrast in the image. Pixels closer to the sensor have brighter values (e.g., the corner of the box). The rest of the points have "darker" grey values, meaning that they are farther away from the 3D sensor. The exact location of the corner is found by applying a local maxima operator to the example image in Figure 20(e).

Once the corner point is obtained, its 3D coordinate description can be computed. Figure 20(e) shows the location of the point P_a, which corresponds to the corner which is closer to the Kinect sensor. The point P_a can then be mapped to the world frame $\{W\}$ (as mentioned before, for the general application of the system as shown in Figure 1, the world coordinate can be selected to be located on the wearable headgear module). As shown in Figure 21, a transformation ${}^{W}_{K}T$ can be used to find the coordinates of P_a, with respect to the frame $\{W\}$, using the following equation: ${}^{W}P_a = {}^{W}_{K}T\,{}^{W}P_a$. From the position vector ${}^{W}P_a$ the "z" coordinate has the height of the sample object with respect to $\{W\}$.

The last step for the object description is defining the object's top surface. This is accomplished by mapping the top edges of the object to the world coordinate frame. A set of points which have similar height to the main corner point is used to define the top surface of the object. In Figure 20(f) the height in world coordinates of the corner is computed via the transformation ${}^{W}_{K}T$. Figure 20(f) shows orange points enclosing the edges of the top surface.

Any point on the surface of the sample object can be entirely described in space with respect to the Kinect frame using the information in the image.



(a) (b) (c) (d) (e) (f)

Figure 20: Object description algorithm. The color image (a), presented to illustrate one sample object, is not part of the processing. The mask of connected components (b) obtained in the former stage is the initial input of this block used. The depth pixels corresponding to each object are segmented using the mask (c). Histogram equalization (d) is used to find the brightest pixel in the image (e). The brightest pixels in the object correspond to the corner (or edge) closer to the Kinect-I. This corner is used to describe the height of the object. The contour in (c) is filtered to find other points with a similar height to the corner (f). The "corner" point and the points with similar height are used to describe the top shape of the sample.

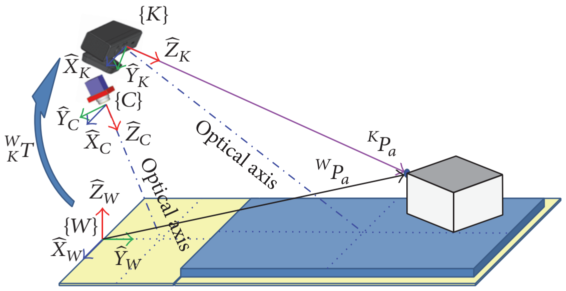


Figure 21: The coordinates of a point in the sample can be described with respect to the frame of the Kinect $\{K\}$. Additionally, using transformation ^W_KT, this point can be described with respect to the world frame $\{W\}$.

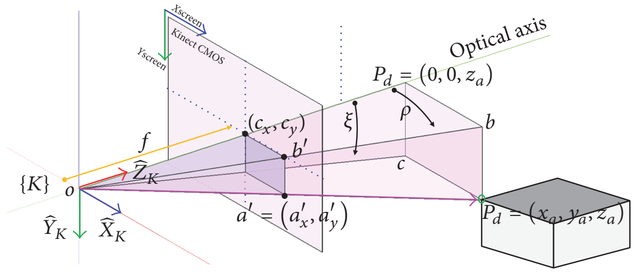


Figure 22: Pinhole camera model for the Kinect-I sensor. The coordinates of a test object can be obtained from the pinhole model using simple trigonometric relations.

Point ^WP_a can be expressed as

$$^WP_a = \begin{pmatrix} x_a \\ y_a \\ z_a \end{pmatrix} = \begin{pmatrix} z_a \tan\rho \\ z_a \tan\xi \\ z_a \end{pmatrix}, \qquad (8)$$

where $\rho = \tan^{-1}((a'_x - c_x)/f)$ and $\xi = \tan^{-1}((a'_y - c_y)/f)$. Figure 22 shows the geometrical diagram which was used to extract these equations using similar triangles as the pinhole camera model. Once the position vector is entirely defined for the Kinect-I frame, the transformation ^W_KT (defined before) can be used to describe the point with respect to the world frame.

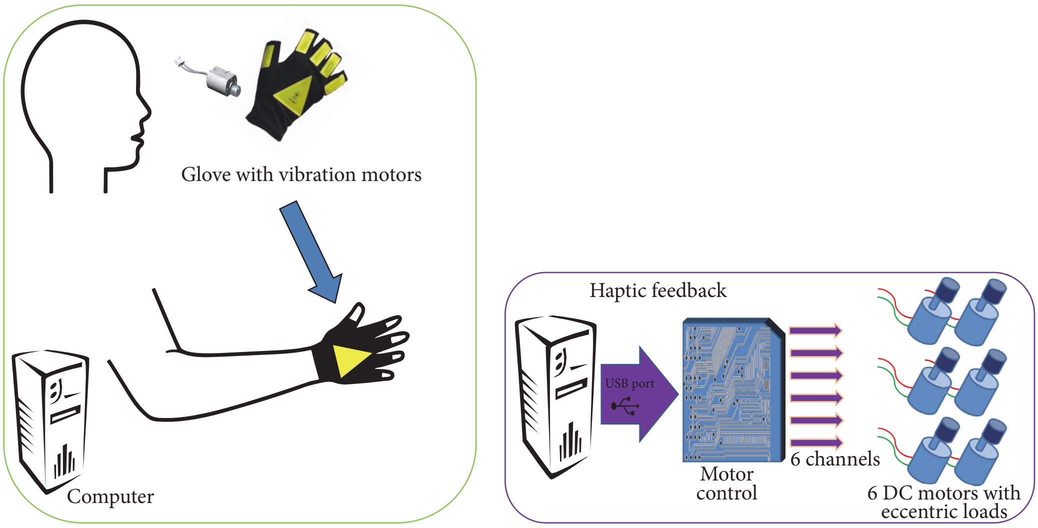
6. Design of a Haptic Glove

There exist various alternatives for introducing haptic feedback. For example, a design of exoskeleton mechanisms was proposed in [30]. This design allows the continuous feedback of the contact forces to the tip of the fingers. In this study and similar to other technologies such as [31], we have proposed a haptic glove with the vibratory actuators for creation of the sense of touch. The haptic glove was designed to provide feedback to the user by activating vibrating elements (tactors) placed on the glove. For the current design six motors were attached to the glove. A USB connection with the PC provides the input for a microcontroller board with 6 PWM outputs to set the vibration frequency of each motor. Figure 23 shows the communication architecture for the motor control.

Considering a simple kinematic model of the hand, it is represented using 6 elements (5 fingers and the dorsal aspect of the hand). Being consistent with such hand description, the 6 tactors are related to these elements and located on top of each finger and on top of the hand. Figure 24 presents the motors attached to the proximal phalanges (on the glove) and the center of the back of the hand.

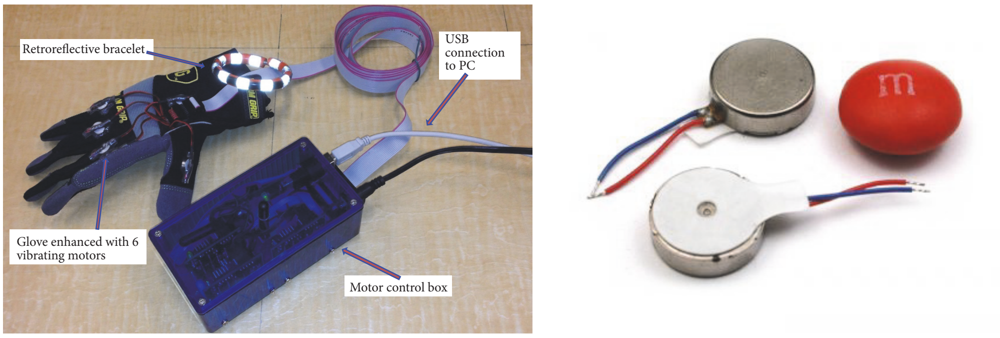
7. Sensor System Integration

With the data provided by the 3D and 2D subsystems, it is possible to construct a map of haptic stimuli. This map of vibrotactile sensations is presented to the user as feedback data with the activation of the tactors on the glove when the user moves his/her hand in a scanning motion along



(a) (b)

Figure 23: Haptic glove concept (a) and haptic glove control architecture (b). Feedback signals are sent to a six-channel motor control box. The frequency and vibration rhythm can be independently controlled for each motor.



(a) (b)

Figure 24: Hardware elements used in the construction of the control box for the glove actuators. In (a) electronic boards from the MIROHOT platform have been repackaged and reused to build the control box for the glove. In (b) the assembled control unit with 5VDC power and connections to the glove actuators.

the haptic region. Data measured from objects placed on the sample platform can be encoded as tactile stimuli on the haptic display area. The haptic map can be represented as a volumetric set created by the intersection of the virtual objects set with the hand location set.

In the haptic map $H_{\text{map}} = A \cap B$, where A refers to the union of all the test volumes representations, combining position, and dimensions of every axis and B refers to the volume occupied by the hand. Every time such an intersection in inhabited, there will be a haptic message (active actuators). In case the intersection is empty, the actuators are inactive.

Figure 25 illustrates the process of building the haptic map. The position and dimensions of sample volumes on the table (b) are computed by the system using the depth image. Models created to represent the objects are virtually represented in the workspace of the user (a). A graphic representation with top and side views is displayed on a monitor ((c), (d)). The gesture and bracelet modules provide the system with the position of the hand. If the hand of the user is out of the volumetric representation, (c) the motors on the glove are set to off. Figure 25(e) illustrates what happens in case the hand of the user intersects the volumetric representation. In this condition, (g) motors are activated. The haptic perception provides the user with information about the position and dimensions of the objects on the table.

The system communicates its output via the haptic display (haptic glove) and a graphical interface. In Figure 26 it can be observed how a test sample is represented in dimensions and location using an orthogonal representation. Figure 26(c) shows the model of a box from the top view (X-Y plane). A side view of the box is in Figure 26(c) using the projection on the Z-Y plane. The projection on Z-X plane is in Figure 26(d). Additionally, the virtual representation of the samples can be seen in Figure 26((b) and (b$'$)) in orange. The mentioned scale factor results in "deformation" in the haptic information. The rectangle in Figure 26(c) is mapped as a trapezoid in

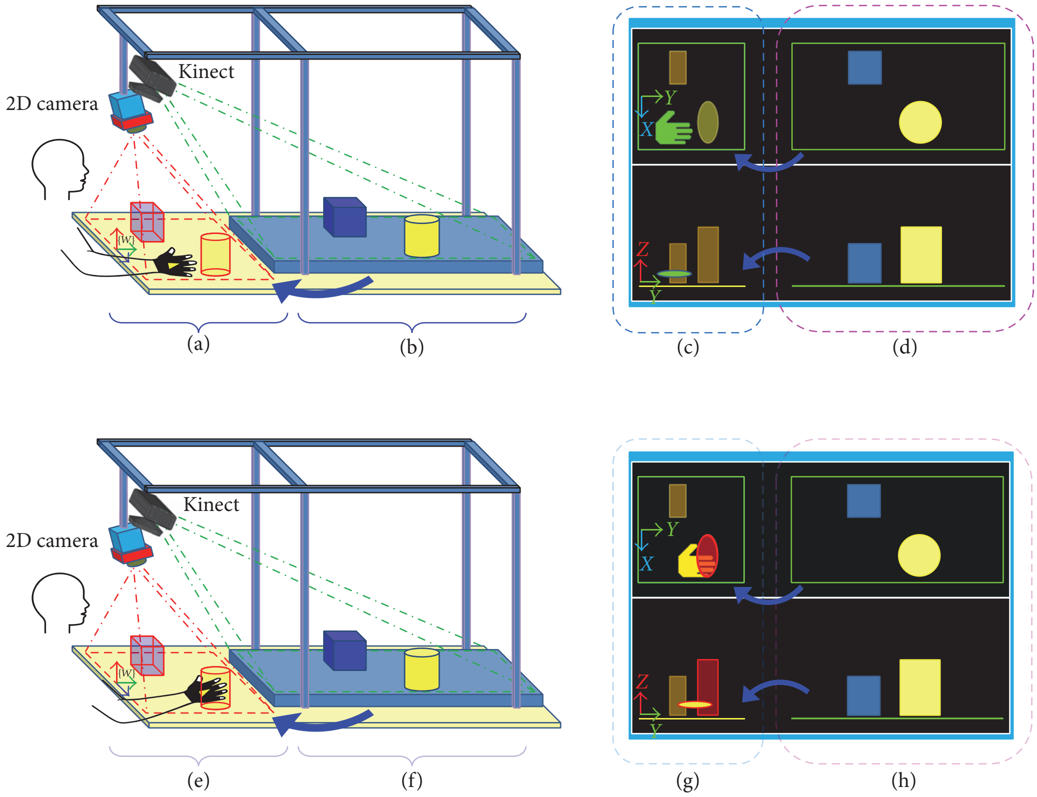


Figure 25: Generation of the haptic signal. Objects on the test table ((b), (f)) are described and modeled with the depth system. They are represented on a user screen with top and side view ((d), (h)). Virtual objects are represented in graphic form in (c) and in tactile form in the tactile region (a). When the hand of the user touches or penetrates the virtual volumes (e) a vibration message is sent to the tactors on the glove.

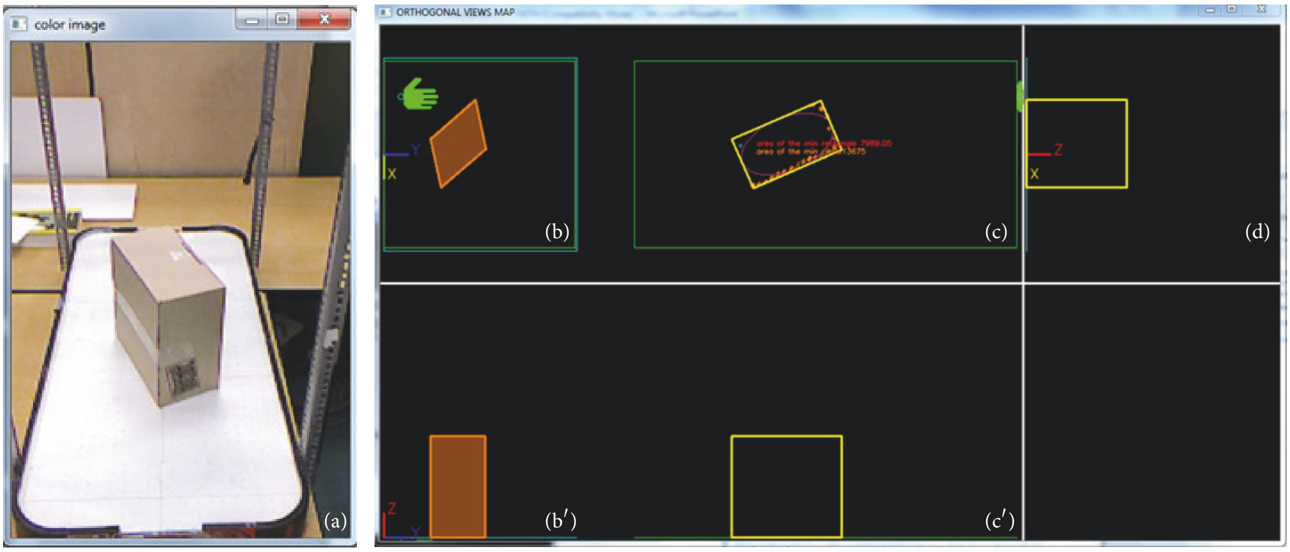

color image
ORTHOGONAL VIEWS MAP
Y
X
(b)
area of the min rectangle 7989.05
area of the min circle 13675
(c)
Z
X
(d)
Z
Y
(b′)
(c′)
(a)

Figure 26: Example of the graphical display of the complete system. A box (a) is represented with a top (c) and 2 side views ((d), (c′)) in the orthogonal projection. A scaled model of the box is represented with top and side view ((b), (b′)). A model of the hand represents the position of the hand in (b) in green.

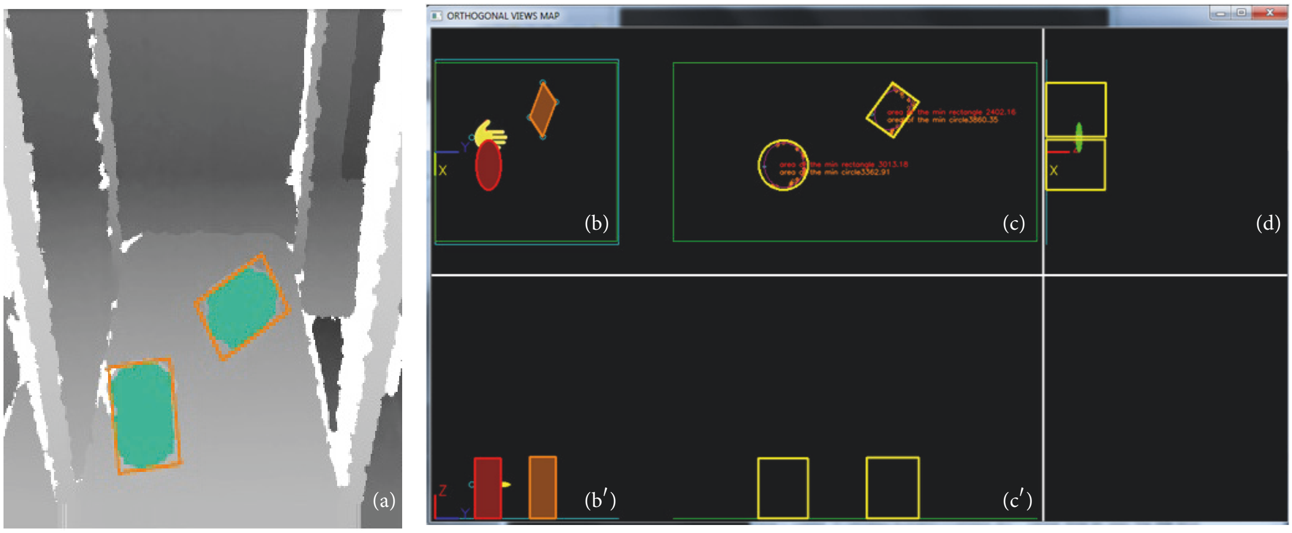

ORTHOGONAL VIEWS MAP
X
Y
area of the min rectangle 2402.16
area of the min circle3860.35
area of the min rectangle 3013.18
area of the min circle3362.91
X
Z
Y
(a)
(b)
(c)
(d)
(b′)
(c′)

Figure 27: Two objects in the depth image (a) are mapped in the haptic display region ((b), (b′)). In this case, the hand of the user is penetrating the volume corresponding to the cylinder, receiving the tactile message.

Figure 26(b). Height and width are displayed without change. In this particular example, the hand of the user is out of the haptic volume.

In a second example (Figure 27), the hand of the user penetrated the haptic representations of a cylinder (in red) in Figure 27(b). Once more the effect of the scale factor is evident in the mapping of the surface of the cylinder in Figure 27(c) onto an oval shape in Figure 27(b). To activate the tactors the two conditions required are the intersection of the hand model and top silhouettes (Figure 27(b)) and that the height of the hand is less than the height of the model (Figure 27(b′)).

Figure 28 shows an example of a user interacting with our proposed multimodal haptic interaction environment. On the right side the graphical interface presents models of the samples on the table. In the center, the haptic display provides cues of the position, orientation, and dimensions of the samples. From the tactile feedback the user can "feel" differences in position, height, width, and length.

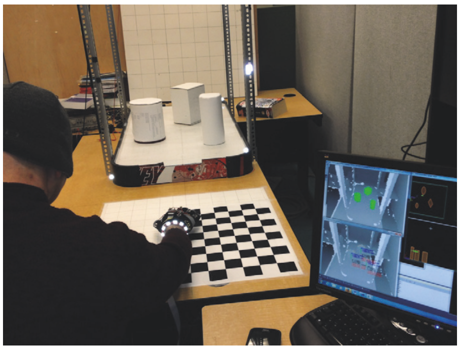


Figure 28: A user is interacting with the proposed multimodal haptic interaction environment. The environment allows the user to perceive the position and dimensions of different objects using haptic feedback.

8. Preliminary User Study

This section presents a user study of our proposed multimodal sensing system for haptic interaction. A group of 11 volunteer users of different genders and age groups ranging from 9 to 50 years old (average age = 29.27 years) were selected to evaluate the performance of our system in controlled conditions. The individuals had no previous training using the developed platform. For this study, four tasks were developed with the goal of measuring the ability of the users to solve a number of requirements based only on the haptic feedback signal.

As part of the tasks the users were required to locate the relative position and distinguish a number of different objects placed on the test table. They were also asked to find the minimum noticeable threshold distance between the objects, identify the shape, and finally distinguish between several geometrical features of objects such as height, width, and length.

8.1. Procedure. Users were first oriented with the general experimental setup. They were then seated in front of the haptic area and also instrumented with the haptic glove and bracelet locator (Figure 29). In order to evaluate aspects related to the haptic feedback, the users were blindfolded during the experiments. The study was divided into four activities. Test objects were randomly selected from a pool of 12 different objects (4 cylinders and 8 boxes with different dimensions). The actuators of the haptic glove were controlled by a 3 V signal which at this voltage can generate an unbalanced angular velocity of 12.000 rpm.

8.1.1. Task 1: Spatial Resolution Experiment. Two wooden test blocks were placed close to each other on the test table. Users were required to locate the blocks and count the

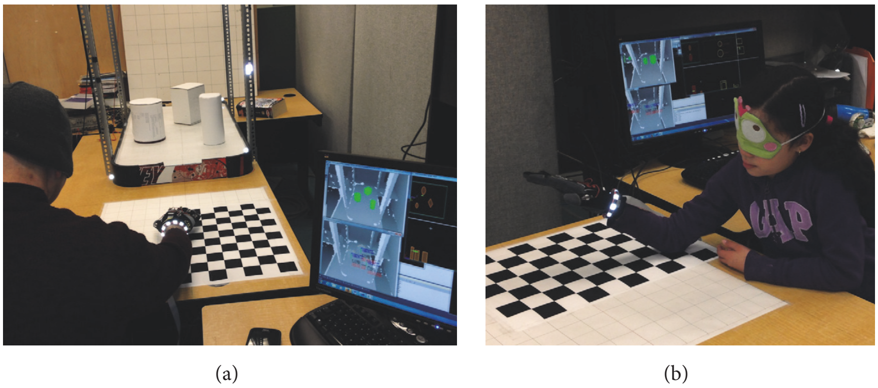
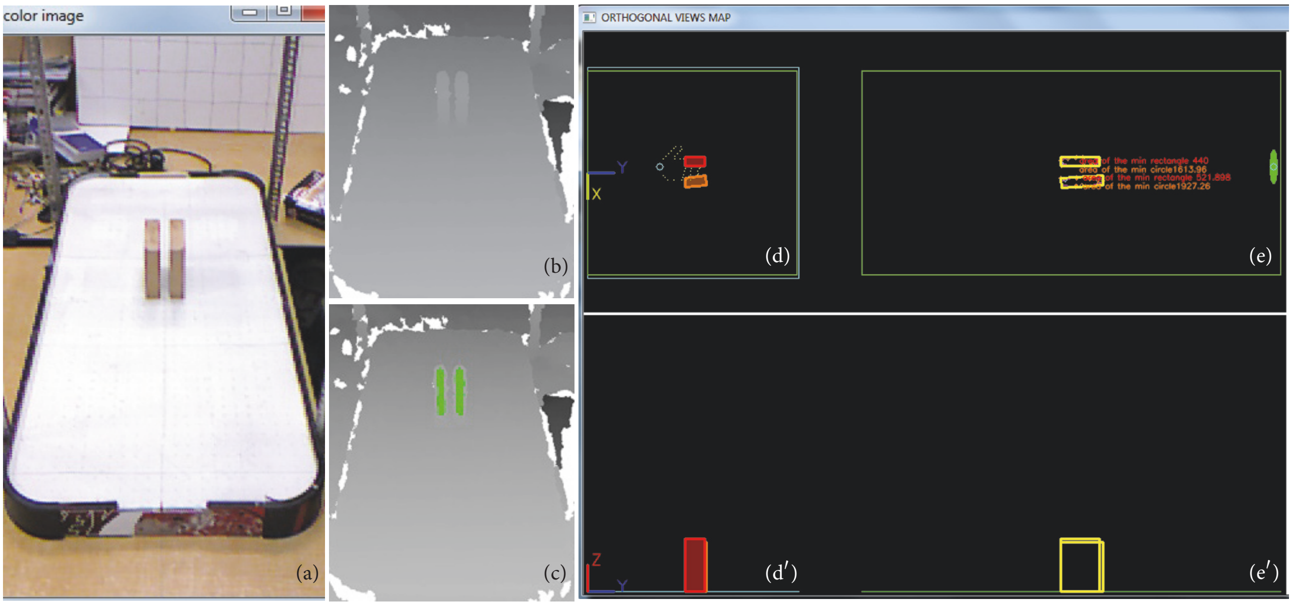


FIGURE 29: Blindfolded users performing different tasks. In (a) a user is locating and counting the number of objects on the test platform. In (b) a young user is able to recognize differences in dimensions and volume just by tactile feedback.



area of the min rectangle 440
area of the min circle1613.96
area of the min rectangle 521.898
area of the min circle1927.26

FIGURE 30: Task 1. (a) Separated blocks are perceived as 2 objects by the depth system, (b), (c), and (d). The user is interacting with one of the blocks (red rectangle in (d)) using one finger.

number of blocks that they could perceive through haptic perception. In consecutive steps, distance between the blocks was reduced in steps of 1 mm until the users were not able to distinguish any noticeable separation between the two blocks. This distance was then recorded as the minimum threshold distance required between the objects in order to distinguish the objects individually. Figures 30 and 31 show screenshots taken during Task 1. In Figure 30 the blocks are separated in the initial stages of the test. In Figure 31 blocks are perceived as a unique object.

8.1.2. Task 2: Spatial Location and Object Counting. A number of objects were randomly selected and placed on the test platform with more than 15 cm separation distance. Blindfolded users were required to locate and count the number of objects in their reachable workspace (Figure 32). Users had no previous knowledge of the number of objects placed on the table for each iteration. The time the users required to scan the whole haptic volume was registered. The recording of time was terminated when users indicated they finished the volume scanning process.

8.1.3. Task 3: Shape-Based Recognition. A cylinder and a parallelepiped block were placed on the platform. Users were required to distinguish differences between the objects and recognize their shapes. Figure 33 shows one of the users using hand gestures to perform the test.

8.1.4. Task 4: Size-Based Discrimination. Users were required to recognize the differences between two objects placed on the platform. The Task was composed of 4 subtasks starting with height, width, length differentiation, and, then, general volume differentiation. For each subtask a new pair of objects was tested. Figure 34 shows an example of one user measuring the sizes of 2 cylinders.

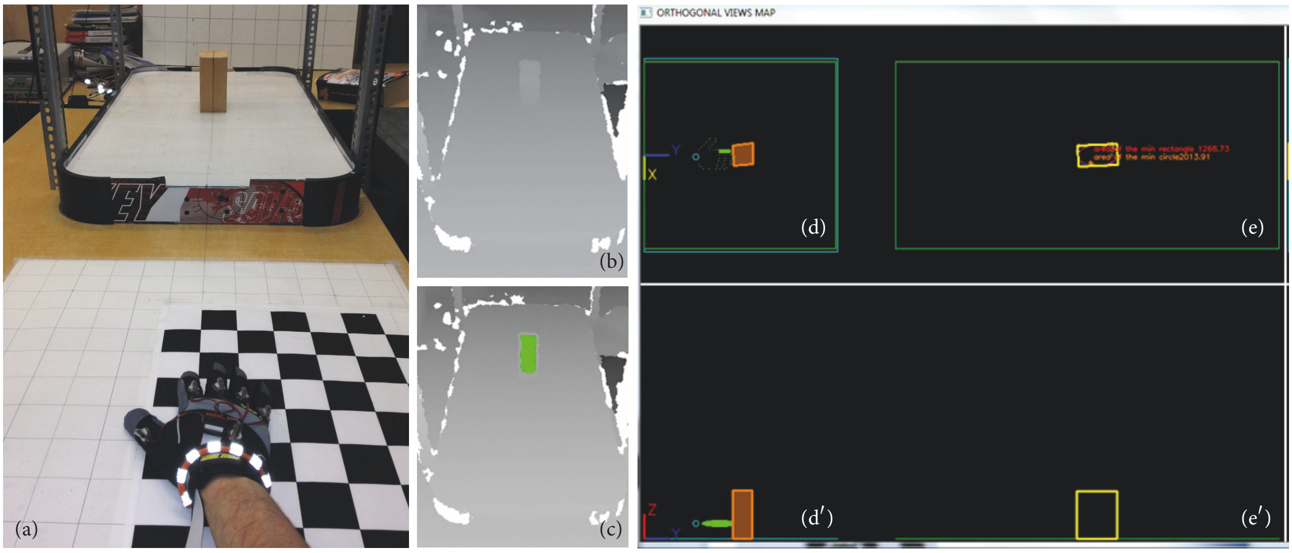

ORTHOGONAL VIEWS MAP
Y
X
area of the min rectangle 1268.73
area of the min circle2013.91
Z
Y
(a)
(b)
(c)
(d)
(e)
(d′)
(e′)

FIGURE 31: Task 1: when the distance between the blocks is under a minimum threshold they are perceived as one block by the user via tactile feedback. In (a) the 2 blocks are side by side. They are represented in (e) and (e′) as a unique block.

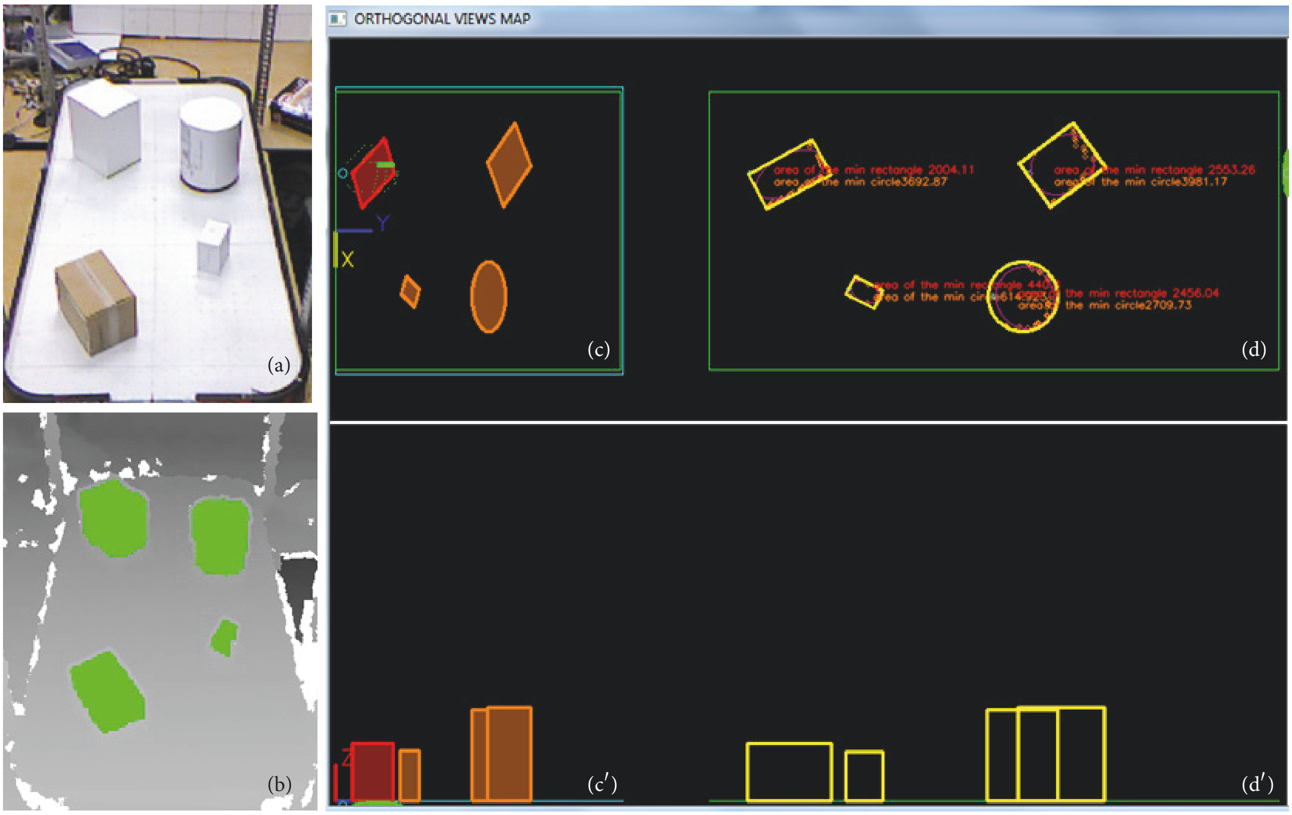

ORTHOGONAL VIEWS MAP
Y
X
Z
area of the min rectangle 2004.11
area of the min circle3692.87
area of the min rectangle 2553.26
area of the min circle3981.17
area of the min rectangle 440
area of the min circle
area of the min rectangle 2456.04
area of the min circle2709.73
(a)
(b)
(c)
(d)
(c′)
(d′)

FIGURE 32: Task 2: during the test, a number of objects were placed on the table to be located and counted by the user. The user is interacting with the brown box in (a). The interaction is registered and labeled in red in (c) and (c′).

8.2. Results. Task 1: it was found that the average minimum threshold distance between objects to be identified as two distinct objects was 19.55 mm (with stdev = 3.11 mm). This distance was measured by having two identical blocks beside each other. The average was computed from the responses of the volunteers.

Task 2: the average time to complete this task was 58 seconds (stdev = 30 s). A total of 110 objects were placed on

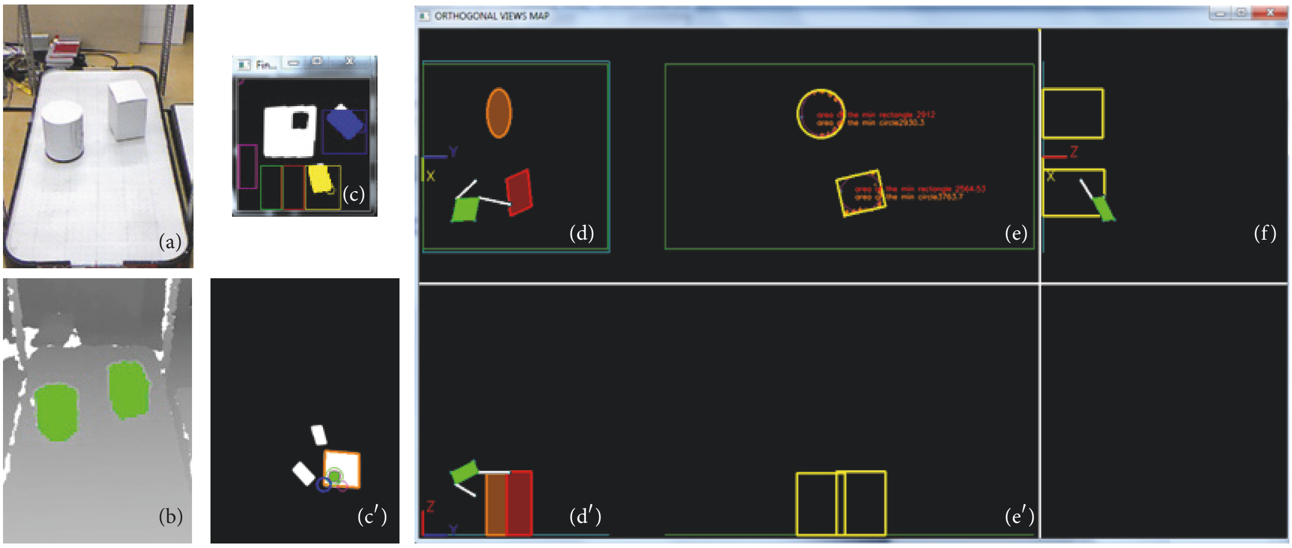

(a)
(b)
Fin...
(c)
(c′)
ORTHOGONAL VIEWS MAP
Y
X
(d)
area of the min rectangle 2912
area of the min circle2930.3
area of the min rectangle 2564.53
area of the min circle3763.7
(e)
Z
X
(f)
Z
Y
(d′)
(e′)

FIGURE 33: Task 3: two samples with different shapes and similar heights are presented to the user to be identified (a). Using one finger gesture ((c), (c′)), the user scans the top silhouettes. In (d) and (d′) the user is interacting with the parallelepiped volume.

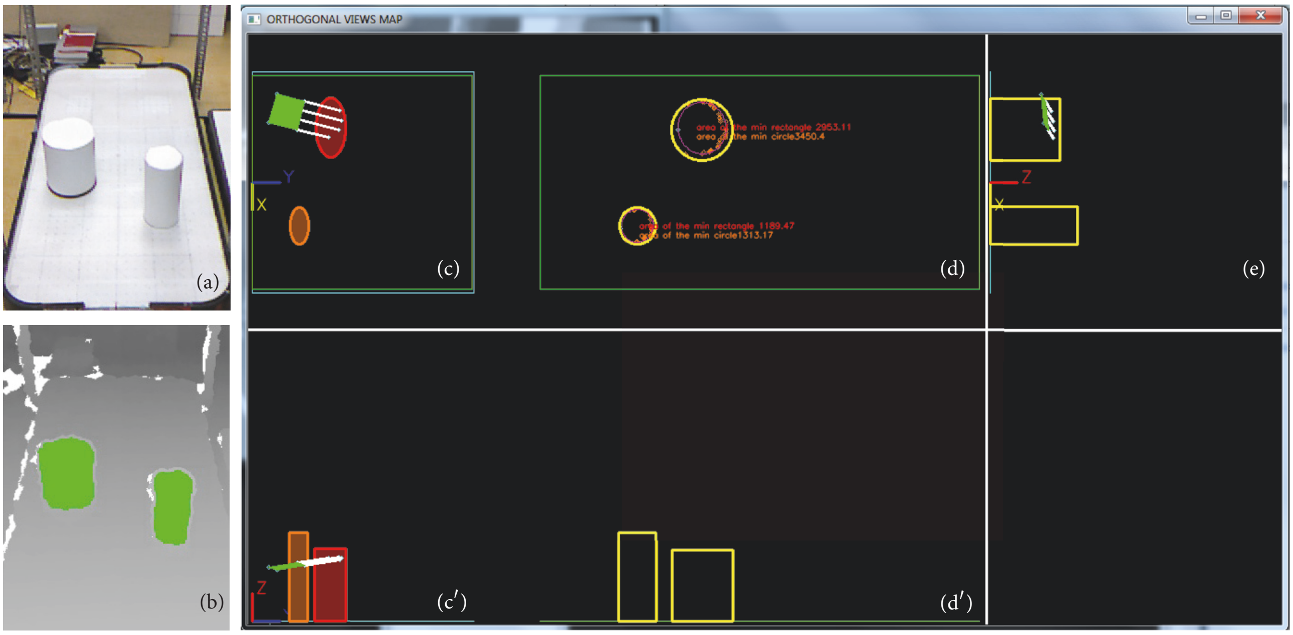

(a)
(b)
ORTHOGONAL VIEWS MAP
Y
X
(c)
area of the min rectangle 2953.11
area of the min circle3450.4
area of the min rectangle 1189.47
area of the min circle1313.17
(d)
Z
X
(e)
Z
(c′)
(d′)

FIGURE 34: Task 4: cylinders of different dimensions are presented to the user for recognition ((a), (b)). User is interacting with the haptic representation of the cylinder with bigger diameter ((c), (c′)) (in red).

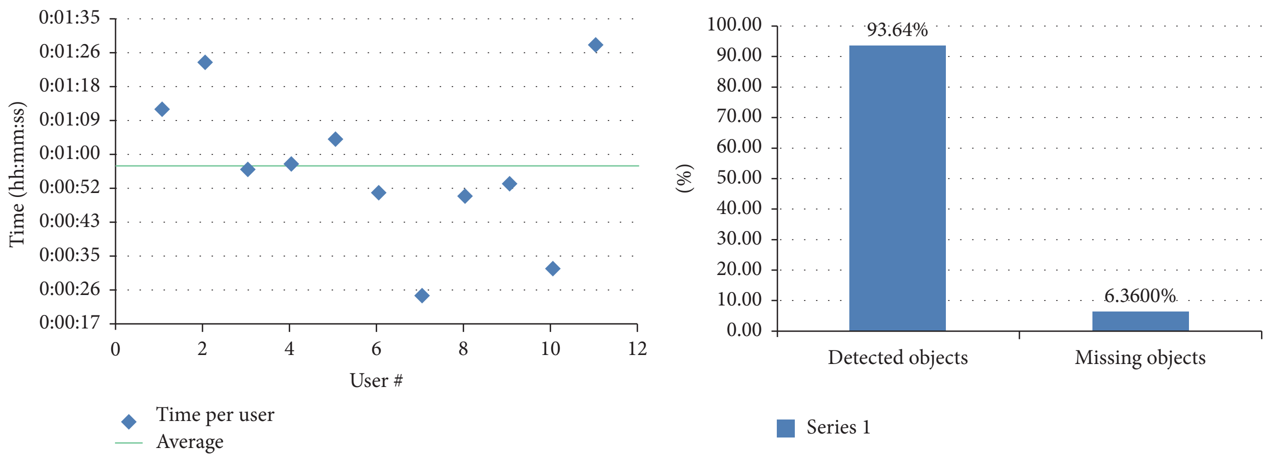
the table for all the users. From this total, 103 objects were properly located and detected resulting in a success rate of 93.64%. On the other hand, in 2 cases, objects were counted several times by the same user resulting in a false detection rate of 1.82% (Figure 35).

Task 3: only 3 users from the 11 were able to properly identify the objects based on their shape. Four users (27.27%) identified the cylindrical containers as boxes and vice versa. The 4 remaining users reported not being able to distinguish any difference between the samples. In total 8 users failed the recognition test (72.72%) (see Figure 36).

Task 4: a total of 44 object pairs were compared based on height, width, length, and total volume. All users were able to find the difference between object pairs with 100% efficiency for all subtasks (Figure 37).

9. Discussions and Conclusions

This paper presents a multimodal sensor interface system for haptic interaction. The developed interface uses depth images to model a set of real objects present in the scene. An image processing module estimates the position orientation



(a)

(b)

Figure 35: Results of Task 2: (a) time required per user to complete the Task and (b) 93.64% of the total of objects presented to the users were detected properly.

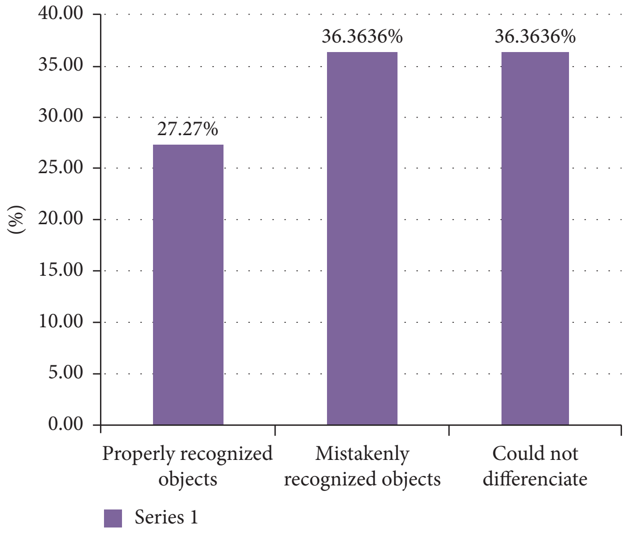


Figure 36: Results of Task 3: 27.27% of the users identified the objects properly. 72.72% failed in the shape identification.

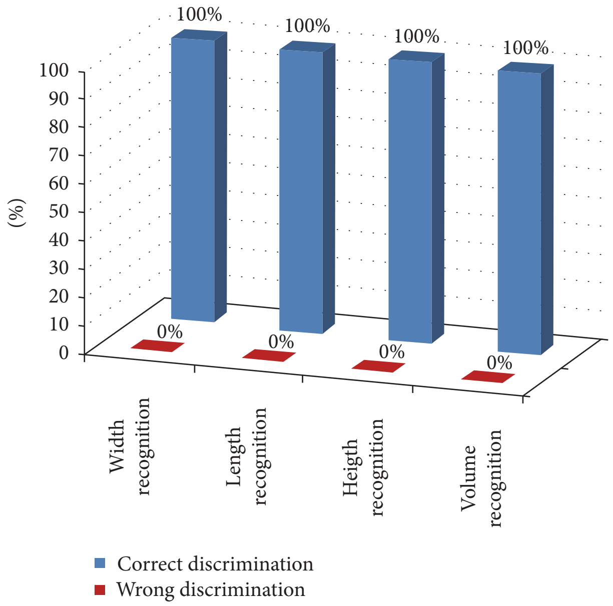


Figure 37: Results of Task 4: in all the subtasks users were able to differentiate between samples with bigger and smaller dimensions.

and configuration of the hand of the user to provide haptic stimuli when the user interacts with virtual objects. The user can perceive the relative position and dimensions of remote objects in the scene scanning with a vibrotactile glove inside the haptic display.

The bracelet location module provides an estimation of the position of the wrist of the user with an error in position of 22.07 mm. When the error in position is contrasted with the dimensions of the haptic interaction area (cube of 600 mm on each side) the error can be expressed as a percentage (22.07 mm/600 mm = 3.66%).

On the other hand, the gesture detection module provided a simplified representation of the user's hand. A kinematic model for the hand was proposed based on the relative position of reflective segments around a core pattern associated with the back of the hand. To obtain information about the orientation (rotation matrix) and location of the hand (position vector), perspective transformations are used. Making use of known patterns, it was possible to build a system to recover 3D information from a scene having as an input a 2-dimensional image. As a result, a 6-DOF description of the hand was obtained.

Recovering pitch, roll, and yaw angles associated with the hand frame description from perspective transformed images imply more accuracy in the determination of the rotation about the Z axis (yaw) and less accuracy in pitch and roll angles. This effect was evident in the experiments performed to obtain the error in these angles (e.g., average error in yaw angle $\alpha = 3.37^\circ$). The roll and pitch average errors were 5.59° and 7.80°, respectively. It was also determined that the glove design was able to provide position coordinates for the hand with a similar precision that was obtained for the bracelet system. Error in position using only perspective transformation matrix for the glove was 23.21 mm.

One of the goals of the optical and lighting technique used for both the bracelet and glove modules was to simplify the image processing stages required to do adequate segmentations of the elements of interest. The combination of reflective markers, coaxial IR illumination, and IR filters and the synchronization of short strobe pulses with shutter time of the camera resulted in images with very high contrast that virtually eliminated the necessity of resource-consuming vision algorithms to obtain the foreground elements from the image. The proposed technique allows the capture system to tune to the target (glove) providing its own illumination. This approach proved to offer stable results under cluttered backgrounds and uncontrolled light conditions (ambient light). It is important to mention some of the limitations of the use of reflective surfaces in the proposed configurations.

The use of the bracelet and the marker-enhanced glove allows the system to obtain a more accurate estimation of the position information of the hand in different situations. The use of both location mechanisms can be redundant when the hand pattern is visible, yet in some hand orientations, the coordinates from the hand pattern are not available due to extreme angles of rotation (of the hand). In these cases, the location data can be obtained from the bracelet. The bracelet was designed to be visible even under extreme wrist rotations, providing a reliable method to estimate the position of the hand in those cases when position cannot be computed from the hand pattern.

As part of the depth image processing system, a complete module was designed to transform depth data obtained from 3D images into models of the objects under examination. In this module, the regions corresponding to the samples were segmented by a combination of stages using image arithmetic, background subtraction, and a set of morphology and connected components filters to obtain the regions of interest. Some descriptors were used to build models of the samples from a set of simple geometrical volumes. The modeled objects were used to build a virtual haptic representation. The accuracy of the depth system to locate and describe objects was measured. The average error in the location of a set of samples around the workspace was 9.99 mm on the X axis, 13.84 mm on the Y, and 5.08 mm on the Z axis of $\{W\}$ (global coordinates). The small values of error are the result of the calibration process used to reduce the effect of distortion in the depth images and adequate estimation of the transformation relating the position of the elements with respect to the world frame and the model-based representation used for the samples.

The performance of three of the four modules was measured individually. When contrasting the error values results in location of a point in 3D space, the smaller error is related to the depth imaging module with an error in location of 19.58 mm (standard deviation = 8.20 mm). The bracelet location error was 22.07 mm in position (standard deviation = 11.55 mm) and the glove location error was 23.21 mm (standard deviation = 16.06 mm). The error values can be considered in a similar range with small variations that can be related to the accuracy of the camera calibration values obtained for both subsystems. Additionally, it is important to consider that error values are functions dependant on the work distance. This distance, in the case of the 2D camera, is around 1 meter (distance between the camera and the center of the workspace). In the case of the 3D camera the distance from the sensor to the center of the test table is 1.7 m. With this consideration, the error related to the depth camera is a smaller proportion of the workspace (19.58 mm/1700 mm = 1.15%) compared with the error of the glove as a proportion of the 2D camera workspace (23.21 mm/1000 mm = 2.23%).

Recognition by height, width, length, and volume was 100% effective. All users could classify objects pairs considering these parameters. Future improvements on the platform include a 6-DOF model of the hand to increase the accuracy of finger location and the activation of all the tactors on the glove. Additional improvements can be achieved by considering complex volume rendering. In this study volumes are considered to be constant section geometrical elements represented as parallelepiped and cylindrical volumes. More geometrical volumes will be added in future iterations increasing the set of primitives used for object description and rendering.

The results of the experiments with users show that haptic interaction can be used as a tool to provide an effective representation of real objects in a scene. In most of the activities, volunteers were able to locate the relative position of the samples and compare them effectively with respect to several criteria. The recognition of objects based on shape can still be a subject of further improvements to increase the effectiveness of the recognition. The results obtained also suggest that the developed haptic display could also be used to represent synthetic objects as part of a virtual scene.

From [7, 8] it was also anticipated that there will exist a limit in shape recognition based on haptic feedback alone. Additional modalities can be included in enhancing the shape recognition task. Various approaches can be explored as a part of the future work such as (a) study of various scale factor, for example, (b) depth computation of models at after update rate (current rate is set at 15 frames per second which results in considerable amount noise along the boundaries of objects) process, and (c) addition of textures to both inside and outside surfaces of the objects. As such, different haptic frequencies can be associated with different surfaces of an object which could facilitate the contour recognition.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. Turk, “Multimodal human-computer interaction,” in *Real-Time Vision for Human-Computer Interaction*, pp. 269–283, Springer, Berlin, Germany, 2005.

[2] V. Van Wassenhove, K. W. Grant, and D. Poeppel, “Visual speech speeds up the neural processing of auditory speech,”

Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 4, pp. 1181–1186, 2005.

[3] S. Oviatt, P. Cohen, L. Wu et al., “Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions,” *Human-Computer Interaction*, vol. 15, no. 4, pp. 263–322, 2000.

[4] M. Zöllner, S. Huber, H. Jetter, and H. Reiterer, “NAVI—a proof-of-concept of a mobile navigational aid for visually impaired based on the Microsoft kinect,” in *Human-Computer Interaction—INTERACT 2011*, P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque and, and M. Winckler, Eds., Springer, 2011.

[5] V. Filipe, F. Fernandes, H. Fernandes, A. Sousa, H. Paredes, and J. Barroso, “Blind navigation support system based on Microsoft Kinect,” in *Proceedings of the 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-Exclusion (DSAI ’12)*, pp. 94–101, July 2012.

[6] S. L. Hicks, I. Wilson, L. Muhammed, J. Worsfold, S. M. Downes, and C. Kennard, “A depth-based head-mounted visual display to aid navigation in partially sighted individuals,” *PLoS ONE*, vol. 8, no. 7, Article ID e67695, 2013.

[7] V. Khambadkar and E. Folmer, “GIST: a gestural interface for remote nonvisual spatial perception,” in *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST ’13)*, pp. 301–310, St. Andrews, UK, October 2013.

[8] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo, “A usability study of multimodal input in an augmented reality environment,” *Virtual Reality*, vol. 17, no. 4, pp. 293–305, 2013.

[9] V. A. Prisacariu and I. Reid, “3D hand tracking for human computer interaction,” *Image and Vision Computing*, vol. 30, no. 3, pp. 236–250, 2012.

[10] R. Y. Wang and J. Popović, “Real-time hand-tracking with a color glove,” in *Proceedings of the ACM Transaction on Graphics (SIGGRAPH ’09)*, vol. 28, no. 3, New Orleans, La, USA, August 2009.

[11] C. Diaz and S. Payandeh, “Toward haptic perception of objects in a visual and depth guided navigation,” in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC ’14)*, pp. 3470–3475, October 2014.

[12] C. Diaz and S. Payandeh, “Preliminary experimental study of marker-based hand gesture recognition system,” *Journal of Automation and Control Engineering*, vol. 2, no. 3, pp. 242–249, 2014.

[13] X. Zhang and S. Payandeh, “Application of visual tracking for robot-assisted laparoscopic surgery,” *Journal of Robotic Systems*, vol. 19, no. 7, pp. 315–328, 2002.

[14] C. Diaz, *An experimental study of remote multi-modal interface in robotic systems [Master of Applied Science Thesis]*, Simon Fraser University, 2014.

[15] Y. Nakazato, M. Kanbara, and N. Yokoya, “Localization of wearable users using invisible retro-reflective markers and an IR camera,” in *Stereoscopic Displays and Virtual Reality Systems XII*, vol. 5664 of *Proceedings of SPIE*, June 2005.

[16] Y. Nota and Y. Kono, “Augmenting real-world objects by detecting ‘invisible’ visual markers,” in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST ’08)*, October 2008.

[17] Y. Nakazato, M. Kanbara, and N. Yokoya, “Localization system for large indoor environments using invisible markers,” in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST ’08)*, pp. 295–296, ACM, Bordeaux, France, October 2008.

[18] P. Asadzadeh, L. Kulik, and E. Tanin, “Gesture recognition using RFID technology,” *Personal and Ubiquitous Computing*, vol. 16, no. 3, pp. 225–234, 2012.

[19] R. Y. Wang and J. Popović, “Real-time hand-tracking with a color glove,” *ACM Transactions on Graphics*, vol. 28, no. 3, article no. 63, 2009.

[20] J. Wang and S. Payandeh, “A study of hand motion/posture recognition in two camera views,” in *Proceedings of the International Symposium on Visual Computing*, pp. 314–323, December 2015.

[21] S. Malik, C. McDonald, and G. Roth, “Hand tracking for interactive pattern-based augmented reality,” in *Proceedings of the ACM 1st Internaltional Symposium on Mixed and Augmented Reality*, pp. 117–127, Darmstadt, Germany, 2002.

[22] R. Hartley and A. Zisserman, *Multi View Goemetry in Computer Vision*, Cambridge University Press, 2003.

[23] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[24] R. Macknojia, A. Chavez-Aragon, P. Payeur, and R. Laganiere, “Calibration of a network of Kinect sensors for robotic inspection over a large workspace,” in *Proceedings of the IEEE Workshop on Robot Vision (WORV ’13)*, pp. 184–190, January 2013.

[25] M. Camplani and L. Salgado, “Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 122–136, 2014.

[26] A.-T. Nghiem and F. Bremond, “Background subtraction in people detection framework for RGB-D cameras,” in *Proceedings of the 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS ’14)*, pp. 241–246, IEEE, Seoul, South Korea, August 2014.

[27] S. Tanimoto, *An Interdisciplinary Introduction to Image Processing*, MIT Press, 2012.

[28] T. Moeslund, *Introduction to Video and Image Processing*, Springer, 2012.

[29] J. J. Koenderink, *Solid shape*, MIT Press Series in Artificial Intelligence, MIT Press, MA, USA, 1990.

[30] Z. Ma and P. Ben-Tzvi, “Design and optimization of a five-finger haptic glove mechanism,” *Journal of Mechanisms and Robotics*, vol. 7, no. 4, Article ID 041008, 8 pages, 2015.

[31] Road to VR, http://www.roadtovr.com/.