Wee Ching Pang
BeingThere Centre
Institute for Media Innovation
Nanyang Technological University

Gerald Seet*
Xiling Yao
Robotics Research Centre
School of Mechanical and
Aerospace Engineering
Nanyang Technological University

A Study on High-Level Autonomous Navigational Behaviors for Telepresence Applications

Abstract

This paper presents a framework enabling navigational autonomy for a mobile platform with application scenarios specifically requiring a humanoid telepresence system. The proposal promises a reduced operator workload and safety during robot motion. In addition, the framework enables the inhabitor (human controlling the platform) to provide inputs for head and arm gesticulation. This allows the inhabitor to focus on interactions at the remote environment, rather than being engrossed in controlling robot navigation. This paper discusses the development of higher-level, human-like navigational behaviors such as following, accompanying, and guiding a person autonomously. A color histogram comparison and position matching algorithm has been developed to track the person using the Kinect sensors. In addition to providing a safe and easy-to-use system, the high-level behaviors are also required to be human-like in that the mobile platform obeys the laws of proxemics and other human interaction norms such as walking speed. This facilitates a higher level of experience for other humans interacting with the robotic platform. An obstacle avoidance function has also been implemented using the virtual potential field method. A preliminary evaluation was also conducted to validate the algorithm and to support the claim of reducing operator cognitive load due to navigation. In general, it was shown that navigation over a given route was accomplished at a faster pace with no instances of collision with the environment.

1 Introduction

Telepresence (Minsky, 1980) is commonly known as a sense of "being there" at an environment that is physically remote from oneself (Sheridan, 1992). Therefore, a telepresence system or application encompasses a set of technologies that enables one to interact effectively with all the sensations and advantages of actually being at the remote site. Through a telepresence system, users can feel each other's presence; this is achieved by capturing, transmitting, and recreating sensual information to the system's users. Pertinent information can include speech, ambient sounds, smell, visual information, and even the actions of remote parties. Collectively, these can bring about a realistic and effective sensation of being there.

Presence, *Vol. 23, No. 2, Spring 2014,* 155–171
doi:10.1162/PRES_a_00178

*Correspondence to mglseet@ntu.edu.sg.

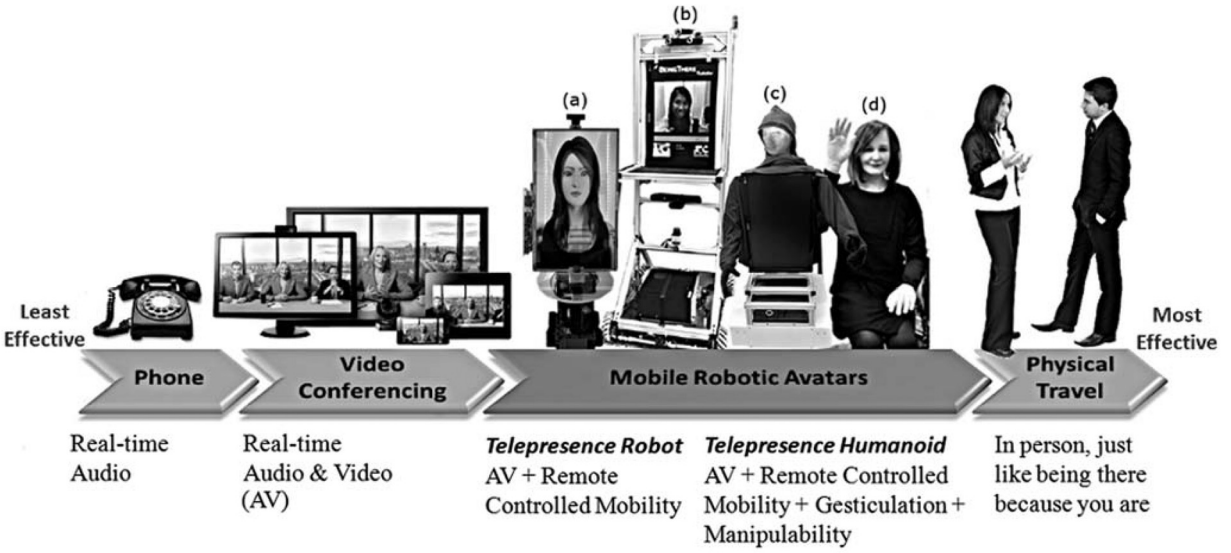

Figure 1. *Spectrum of telepresence applications for telecommunication, including (a) the MAVEN-mini, (b) the MAVEN-I, (c) the MAVEN-II and (d) the Nadine telepresence humanoid.*

1.1 Background of Telepresence for Interpersonal Communication

With the prevalence of the internet and the advances of technology for telepresence, there is a desire to replicate face-to-face interaction such that interaction of a similar degree of richness can take place without the need for the interacting parties to be physically co-located. Such co-location may require international travel, which is time-consuming and contributes to one's carbon footprint. Figure 1 illustrates a spectrum of telepresence technologies designed to support face-to-face interaction between people, especially in terms of natural conversation that involves the exchange of audio and visual information. The spectrum shows how current systems along this spectrum have become, and are continuing to become, increasingly immersive and thus more effective in using telepresence for communication.

On a lower scale of telepresence service is telephony (Walker & Sheppard, 1997), which has vastly increased the range of communication. When a user calls a telephone helpline, service personnel can provide a remedy to a technical issue, even if the technician is miles away. However, telephony typically involves only the transmission of speech between two parties; therefore, richer information such as visual content is absent. Videoconferencing (Turletti & Huitema, 1996) improves on the telepresence experience by enabling visual communication in addition to the exchange of audio, such that geographically separated individuals can also see one another. Telepresence through video-mediated communications ranges from mobile video chatting tools, such as Microsoft Skype, Apple Face Time, and Google Talk, to multiparty video conferences as well as dedicated telepresence boardrooms such as Cisco Telepresence (Szigeti, McMenamy, Saville, & Glowacki, 2009). Video-mediated communications can also include 3D holographic immersive rooms, such as the DVE teleimmersion room (Digital Video Enterprises, 2010). This progression increases the sense of presence and connection between remote participants and can lead to more advanced immersive telepresence systems such as high-fidelity 3D room-based telepresence systems whereby a room can be virtually extended by "joining" remote locations through wall-sized displays.

In the recent past, there has been an emergence of a new communication method where a mobile robot is used to augment communication by integrating

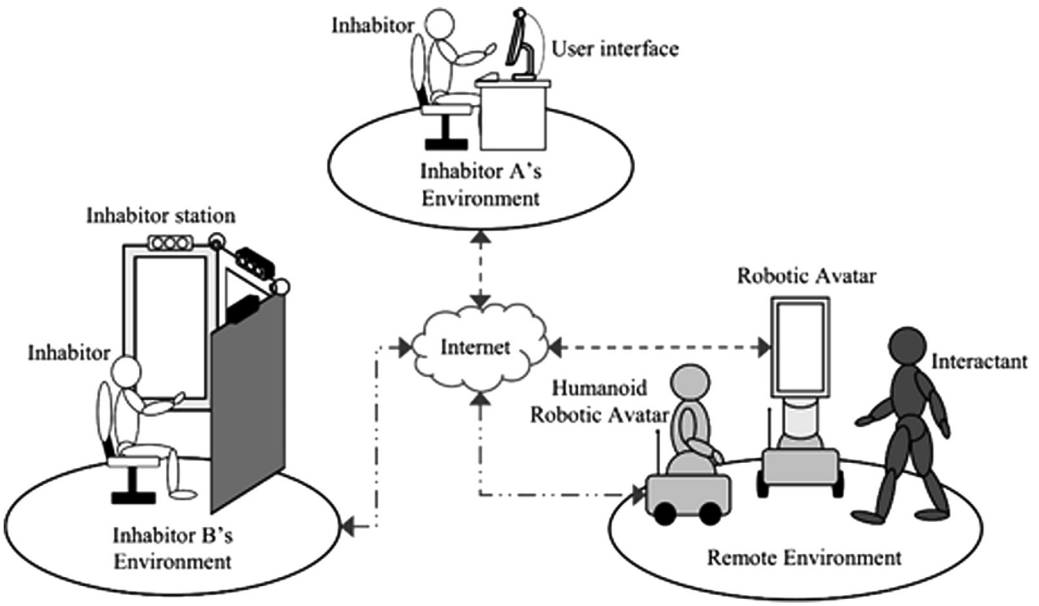

Figure 2. *Elements of a robotic telepresence application.*

video-mediated communication tools with the robot. This method offers the means to connect to a remote location via traditional video conferencing, but also has the added value of moving and actuating in that remote location via the robotic system. Such robot-mediated communication tools are commonly known as telepresence robots, although some may refer to them as remote presence systems (Willow Garage, 2011; InTouch Technologies, 2011), virtual presence systems (Anybots, 2010), embodied social proxy (Venolia et al., 2010), or robotic avatars (Lincoln, Welch, Nashel, Ilie, & Fuchs, 2009; Seet, Pang, & Burhan, 2012; Seet, Pang, Burhan, & Chen, 2012). The use of a robotic avatar as an emerging telepresence application is evidenced by the increasing amount of commercial systems available and the associated research efforts. These robots have been developed to perform in a plethora of applications such as performing medical rounds in healthcare institutions (Thacker, 2005; Ellison et al., 2004) and conducting ad hoc conversations in office environments (Lee & Takayama, 2011).

This paper aims to report on our development work to implement one such robot-mediated telepresence application. This work has been performed primarily at the Nanyang Technological University (NTU). The developed robotic avatar has been named MAVEN, Mobile Avatar for Virtual Engagement by NTU (Seet, Pang, & Burhan, 2012; Seet, Pang, Burhan, Chen et al., 2012). It allows the inhabitor to establish his or her presence with the use of a 2D transparent screen, as shown in Figure 1(a), or a 2D graphical display, as depicted in Figure 1(b), or a 3D physical display, as depicted in Figures 1(c) and (d). This paper proposes a framework to discuss a robot-mediated telepresence application, which uses a humanoid robotic avatar.

2 A Robotic Telepresence System

The elements of a typical robotic telepresence system can be summarized in the framework shown in Figure 2. This paper adopts specific terms to refer to the robotic system and its users. These terms are listed and described in the following paragraphs. An *avatar* is a machine that represents a particular person in a real-world environment. It is a robot if it exhibits a reasonable degree of autonomy. A *humanoid robotic avatar* is thus an anthropomorphic version of an autonomous avatar.

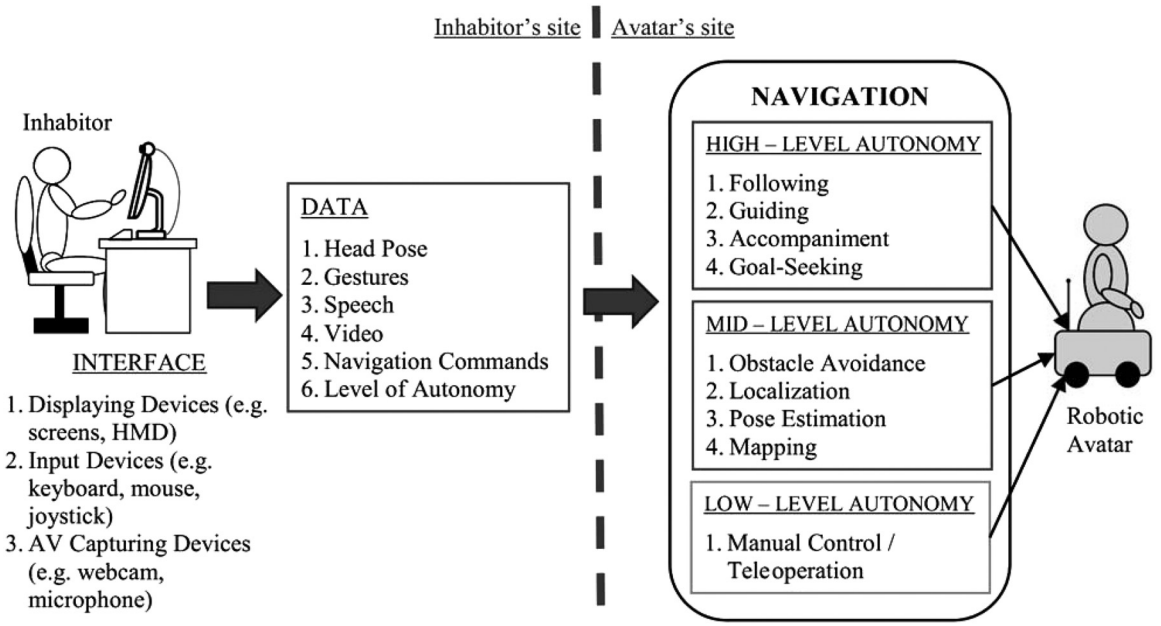

Figure 3. *Framework for a telepresence robot with various modes of navigational autonomy.*

An *inhabitor* is the user who controls the robotic avatar and uses it to represent him or her at the remote location. In the current state-of-the-art, the inhabitor *inhabits* the robotic avatar via a computer interface. It is not considered full *embodiment*, because the robotic avatar cannot be a full manifestation for the inhabitor due to technological limitations in current robotic avatars.

An *interactant* is another type of user. The interactant is someone who interacts with the robotic avatar at the *remote* environment. This remote environment or site is the space where the robot and the interactant are situated. Multiple interactants can be co-located with the robotic avatar.

A *user interface* is the interface that is used to control the robotic avatar as well as to display the information acquired from the robotic avatar. An *inhabitor station* is an advanced and immersive user interface, which is composed of acquisition sensors and feedback systems, for controlling the animatronic robotic avatar.

The use of robots for telecommunication would involve a two-way transmission of various data in real time. Such data can include audio, video, and other contents that are necessary for verbal and nonverbal communication. Figure 3 illustrates a framework that depicts the flow of data from the inhabitor to the mobile robotic avatar at the remote site. While the robotic avatar may be imbued with a multitude of behaviors for various needs such as interaction and object manipulation, this paper pays particular attention to navigational behavior. Figure 3 shows the various modes of navigational autonomy that the mobile robotic avatar is equipped with.

In Figure 3, data from the inhabitor is acquired at the inhabitor station via traditional input devices such as a mouse, keyboard, joystick, and webcam. A graphical user interface facilitates the input of data and also presents information to the inhabitor. Other methods of input and displaying data are also possible. These alternative methods include the use of tablets and immersive large screen display. Data, such as video, speech, and motion commands, are acquired at the inhabitor's station and relayed to the robotic avatar. The video and audio components of this data are displayed at the robotic avatar while the navigational commands are used for robot mobility.

At this time, most commercially available telepresence robots do not have the ability for autonomous

navigation but rely on teleoperation for the robot's navigation. However, such reliance is not ideal and can result in collisions. Experiments conducted by Tsui and her team (Tsui, Desai, Yanco, & Uhlik, 2011) have shown that two-thirds of the participants acting as the inhabitor cause the robot to collide with the environment arranged as an office space. It is thought that collisions occur because the inhabitor lacks situational awareness of the robot's surroundings. This may be due to the high level of cognitive workload experienced by the inhabitor and also the limitation of the graphical user interface for depicting the robot's 3D environment. Cognitive workload can be especially high when the inhabitor is attempting interaction with the interactant while also navigating the robot. Latency effects, typical of current internet communication, can exacerbate the problem of diminished situational awareness.

In addition to navigation, a humanoid avatar can require two additional components for control from the inhabitor: the head (Pang, Burhan, & Seet, 2012) and arm components. A webcam at the inhabitor station can be used to capture data of the inhabitor's head pose and gestures. Head pose and gesture data will then be relayed to the humanoid robotic avatar to control its head and arms. With control inputs needed for these two components in addition to navigation, inhabitor workload is expected to increase yet further.

Due to the high level of workload that the inhabitor is expected to experience, autonomous navigation can be highly beneficial in remotely operating a mobile robot. However, autonomous navigation can include a number of aspects. The robot can navigate such that it not only moves from point to point by itself but can also at the same time maintain a comfortable and safe distance from people who are co-located with the robot. Such attention to achieving a comfortable standoff facilitates natural interaction in addition to facilitating safe and effective autonomous navigation.

To facilitate this kind of interaction via a humanoid robotic avatar, this paper proposes equipping the robotic avatar with adjustable autonomy in navigation to achieve a number of objectives. These objectives include:

1. The reduction of inhabitor cognitive workload;
2. The increase in safety during the navigation of the robotic avatar in a remote environment;
3. The continuous navigation of the robotic avatar during instances of command latency or breaks in instructions pertaining to navigation; and
4. The use of various input methods for the inhabitor such that even handheld devices such as tablets can be employed for use in interaction via a robotic avatar.

If the robot is equipped with adjustable autonomy, then it would be possible for the inhabitor to call upon mid- to lower levels of autonomy should greater control over navigation be required. Such instances can include navigation through cluttered and narrow spaces. However, adjustable autonomy would also allow higher level navigational behaviors to be invoked so that the inhabitor can focus on communicating with interactants rather than on robot navigation.

3 The Necessity for High-Level Navigation Behavior

While we recognize the need for low-level and mid-level navigational autonomy, higher levels of navigational autonomy allow for the inhabitor to also provide inputs for head and arm gesticulation. This is because the inhabitor's arms are freed from the task of manipulating input devices for expressing commands for robot navigation. In addition, the inhabitor can focus on looking at the camera for communication with interactants rather than be engrossed in monitoring robot navigation. As a result, pose data of the inhabitor head becomes meaningful.

This paper discusses the development of these higher-level navigational behaviors. These behaviors include following, accompanying, and guiding an interactant autonomously. Each of these navigation behaviors assist in navigating the robot during a certain type of interaction that is typical between two human individuals, as shown in Figure 4.

In the following behavior, the robotic avatar will be moving behind the person whom it is following.

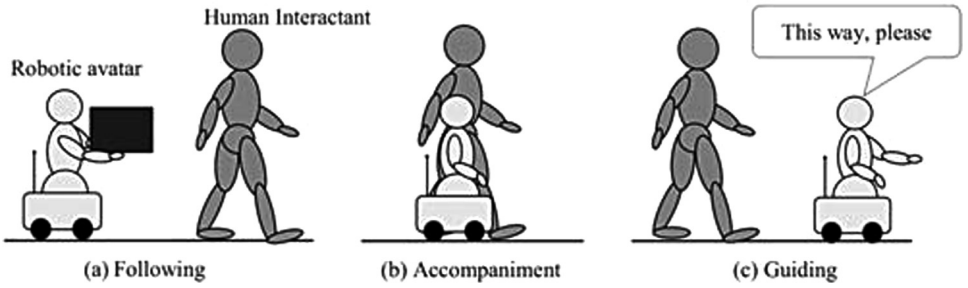

Figure 4. *Different high-level navigational behaviors, including (a) following, (b) accompanying, and (c) guiding a person autonomously.*

For instance, an inhabitor who is using a telepresence robot to visit a museum can choose to use the robot to autonomously follow a human guide. This behavior can also be used to enable the robotic avatar to follow an interactant and provide assistance as shown in Figure 4(a).

In the accompaniment behavior, as seen in Figure 4(b), the telepresence robot will be moving at the side of the interactant whom it is tracking. This is typical of a scenario where two people are having a conversation while walking side by side. In this case, one of the two people in the scenario is replaced with the telepresence robot that is controlled by the remote inhabitor.

In the guiding behavior, as illustrated in Figure 4(c), the robotic avatar will be moving in front of its interactant, such that the inhabitor can provide direction using the telepresence robot to guide the interactant. This behavior is useful when the inhabitor is more familiar with the robot's environment than the interactant.

During these three autonomous behaviors, the robot would try to maintain an appropriate standoff distance from the interactant. The obstacle avoidance capability has been implemented within the system to ensure safe navigation in a human environment. Furthermore, as these behaviors are implemented for a telepresence robot, it is essential for the robot to move at a speed that is similar to that of a real human.

3.1 Related Work

The task of following, accompanying, and guiding an interactant would involve human detection and tracking, which uses the data acquired from a sensor. One method of performing this task is to detect and track a human using a digital color (RGB) camera. Image processing techniques can be applied to the images acquired from the camera to identify blobs that signify a person. Color detection or color histograms (Kwon, Yoon, Park, & Kak, 2005) and feature recognition (Chen & Birchfield, 2007) are some of the other common techniques that have been considered.

The laser range finder is another widely used device for robots to observe the environment. Compared with camera vision data, laser data is more efficient and hence less processing is required (Fod, Howard, & Mataric, 2002). The distance measurements of a laser range finder usually have high accuracy, and the data is not sensitive to ambient noise, such as changing lighting conditions. Therefore, there are many laser-based human detection and tracking studies (Topp & Christensen, 2005; Arras, Mozos, & Burgard, 2007; Gockley, Forlizzi, & Simmons, 2007) that use techniques that process the ranging data to identify the signature of a person's legs. However, in some indoor applications, chairs and tables can be falsely detected as human legs due to the furniture having similar patterns as a human's legs.

Human detection can also be achieved with the use of a depth camera (Loper, Koenig, Chernova, Jones, & Jenkins, 2009), as well as with the new and inexpensive Kinect RGB-D camera (Luber, Spinello, & Arras, 2011). In this paper, the Kinect RGB-D camera is used to detect and track the person to be followed. Although the field of view of the Kinect

sensor is small, it provides sufficient information, such as depth values, audio data, and skeletal mapping, as well as a color image, to perform human detection.

Some of the noteworthy implementations for the back-following behavior are presented in Topp and Christensen (2005), Gockley et al. (2007), Loper et al. (2009), Doisy, Jevtic, Lucet, and Edan (2012), and Cosgun, Florencio, and Christensen (2013). Various works that demonstrate the side-by-side accompaniment behavior include Prassler, Bank, Kluge, and Hagele (2002), Ohya and Munekata (2002), and Morales et al. (2012). Lastly, the work that describes the front guiding behavior includes Montemerlo, Pineau, Roy, Thrun, and Verma (2002), Pacchierotti, Christensen, and Jensfelt (2006), and Burgard et al. (1998).

However, there does not appear to be existing work on combining these behaviors into one system. Moreover, telepresence robots and humans will co-exist in the same space, hence it is important that the robot should be able to move in a manner that is acceptable by the humans around it. There are many works on social navigation that have enabled mobile robots to move from one point to another in the presence of a human (Topp & Christensen, 2005; Gockley et al., 2007; Burgard et al., 1998), using proxemics (Hall, 1990; Kirby, Simmons, & Forlizzi, 2009). However, the navigation system for a telepresence robot should be different from that of a social robot because, unlike a social robot, there is an additional intelligence from the human inhabitor behind a telepresence robot.

3.2 Hardware Configuration

The autonomous following, accompanying, and guiding system is implemented on MAVEN-II. The robot is a holonomic robot with four mecanum wheels. It has an on-board computer for controlling the drive motors. A Fedora operating system was installed as the robot's embedded computer. For this experiment, the maximum forward and lateral speed of the robot was limited to 0.6 m/s, while the rotational speed was limited to 0.9 rad/s.

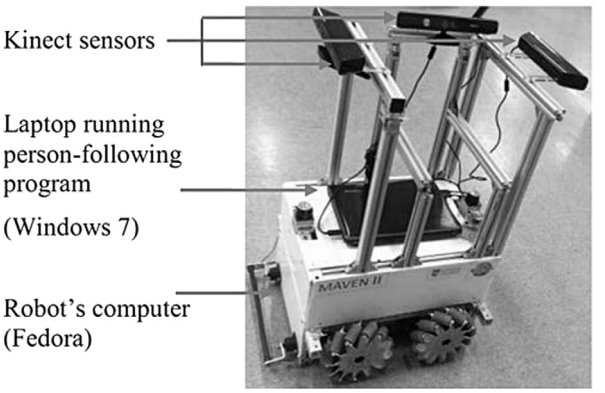

Figure 5. *Hardware configuration of MAVEN-II.*

Currently, the multimodal person-following system was implemented on an additional laptop computer that runs the Microsoft Kinect Software Development Kit (SDK). Three Kinect sensors were mounted on MAVEN-II, as seen in Figure 5; one of them will forward-looking and faces the front of the robot, whereas a second one faces the side of the robot. The third Kinect is mounted at the rear of the robot. These Kinect sensors are used to track the selected interactant and they were connected to the laptop. This laptop is responsible for acquiring data from the Kinect sensors, running the person-following algorithm, and sending velocity and control commands to the embedded computer which in turn controls the robot's movement.

4 Implementation of a System for Autonomous Following, Accompanying, and Guiding

The multimodal person-following system is composed of three main components: human detection and tracking, velocity profiling for each behavior, and obstacle avoidance.

4.1 Human Detection and Tracking Using Kinect and Kinect SDK

The Kinect SDK version 1.5 provided by Microsoft is a set of tools and application programming

interfaces that can be used to create applications by using the Kinect sensor. The Natural User Interface module of the Kinect SDK provides the functionality of accessing the RGB color image data and depth data from the Kinect sensor, and it can be used to detect and track people within the field of view of the Kinect sensor. The Kinect sensor has the capability to detect up to six people in its field of view and it is able to obtain detailed skeletal information, such as the positions and orientations of joints, for a maximum of two people. The Kinect sensor can detect and track people who stand between 0.8 m and 4.0 m from the front of the sensor. Each successfully tracked human can have one of two tracking states: the position only state and the active user tracking state.

In the position only state, only the real-world position (in meters) of the person can be obtained and no other information is available. In the active user tracking state, both the centroid position and the skeletal data, which includes positions and orientations of various joints, are tracked. An ID will be randomly assigned to each detected person and the target can be chosen by selecting the ID of that detected person. In addition to person-tracking, a person-recognition function is required to keep track of the person even when that person is temporarily out of view. An algorithm, which is based on the calculated Euclidean distance between the positions and a color histogram matching technique, has been included to identify and recognize the person. When a new person is detected by the Kinect sensor, after the previously targeted person is temporarily occluded or out of the scene, the newly detected person's current position and HSV color histogram will be compared against the stored position and histogram data. With this comparison, the person with the closest matching position and histogram data will be regarded as the previously tracked person, and the system will automatically resume active tracking.

The position matching measures the Euclidean distance between positions (d_{pos}) as shown in Equation 1, where C_T is the centroid position of the previously tracked person P_T and C_D is the centroid position of the detected person P_D.

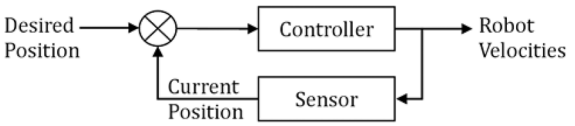

Figure 6. *Feedback control loop for robot speed calculation.*

$$d_{\text{pos}}(P_T, P_D) = \sqrt{(C_{T,x} - C_{D,x})^2 + (C_{T,y} - C_{D,y})^2 + (C_{T,z} - C_{D,z})^2} \quad (1)$$

The histogram-matching technique compares the histogram of the previously tracked person, H_T and the histogram of each detected person, H_D with four types of histogram comparison methods (Bradski & Kaehler, 2008): Correlation, Chi-square, Intersection, and Bhattacharyya. The final histogram-matching distance result is an arithmetic combination of all four histogram distances obtained from these methods as shown in Equation 2.

$$\begin{aligned} d_{\text{hist}}(H_T, H_D) = {} & (1 - d_{\text{correlation}}(H_T, H_D)) \\ & + d_{\text{chisquare}}(H_T, H_D) \\ & + (1 - d_{\text{intersect}}(H_T, H_D)) \\ & + d_{\text{bhattacharyya}}(H_T, H_D). \end{aligned} \quad (2)$$

Subsequently, a score S will be computed for each detected person based on his or her position and histogram matching result, as shown in Equation 3. The detected person with the lowest score is most likely to be the previously tracked person, and the person-following behavior will resume.

$$S = \omega d_{\text{pos}}(P_T, P_D) + \frac{(1 - \omega) d_{\text{hist}}(H_T, H_D)}{(\max(\forall D : d_{\text{hist}}(H_T, H_D))}. \quad (3)$$

4.2 Velocity Profiling

Each behavior within the multimodal person-following system is a proportional feedback control loop. It measures the error between the current robot position and the desired robot position with respect to the interactant. The error is then used to calculate the velocity commands of the robot. The control loop is depicted in the control diagram in Figure 6.

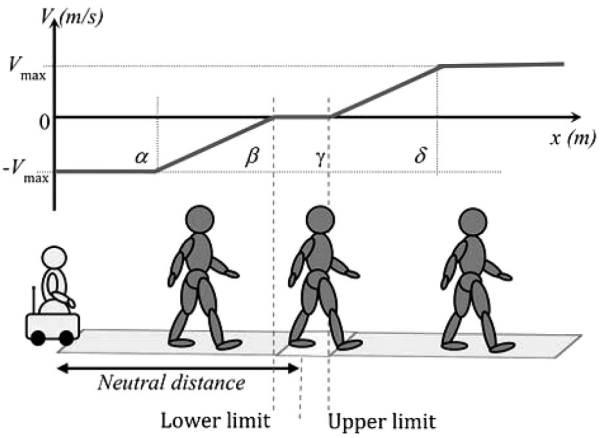

Figure 7. *Rule-based system to determine linear velocity.*

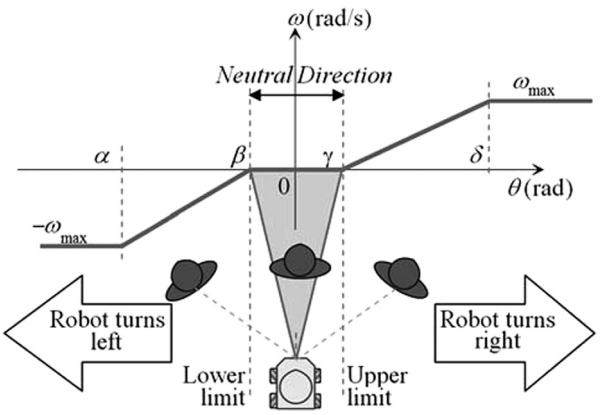

Figure 8. *Rule-based system to determine angular velocity.*

The controller block represents the calculation of the robot's forward, lateral, and rotational velocities. The robot will try to move in a manner such that it maintains a desirable distance from the interactant. It will also adopt an orientation such that the interactant is always at the center of the Kinect's field of view. The desired distance and direction of the interactant relative to the robot is noted as the neutral distance and neutral direction in Figure 7 and Figure 8, respectively, which will be used as the reference for robot speed calculation.

The linear velocity is dependent on the distance between the robotic avatar and interactant, which is denoted as x in Figure 7. The velocity profiling for the linear motion is given by this rule-based formula where we can tune the parameters differently to suit the different following behaviors, as shown in Equation 4.

$$V = V_{\max} f(x) \text{ where } f(x) = \begin{cases} -1 & \text{if } x < \alpha \\ \frac{\beta - x}{\alpha - \beta} & \text{if } \alpha \le x < \beta \\ 0 & \text{if } \beta \le x \le \gamma \\ \frac{x - \gamma}{\delta - \gamma} & \text{if } \gamma < x \le \delta \\ 1 & \text{if } x > \delta \end{cases} \tag{4}$$

Similarly, the angular velocity is dependent on the angular difference between the robotic avatar and the interactant. As shown in Figure 8, if the angular difference is near zero, then the robot will not rotate. If the interactant turns to face another direction, the robot will rotate at a speed that is dependent on the angular difference at that point. The velocity profiling for the angular motion is given by this rule-based formula, where we can tune the parameters differently to suit the different following behaviors, as shown in Equation 5.

$$\omega = \omega_{\max} f(\theta) \text{ where } f(\theta) = \begin{cases} -1 & \text{if } \theta < \alpha \\ \frac{\beta - \theta}{\alpha - \beta} & \text{if } \alpha \le \theta < \beta \\ 0 & \text{if } \beta \le \theta \le \gamma \\ \frac{\theta - \gamma}{\delta - \gamma} & \text{if } \gamma < \theta \le \delta \\ 1 & \text{if } \theta > \delta \end{cases} \tag{5}$$

4.2.1 Following Behavior. In the following behavior, the target interactant remains in front of the robotic avatar. The robot is able to move, by translating and rotating with the interactant while trying to maintain a distance of about 1.2 to 1.5 m. These values were chosen because it is the common social distance between two people during social interactions (Hall, 1990). Therefore, the lower limit and the upper limit of the neutral position have been set to 1.2 and 1.5 m, respectively. The forward translation velocity (tv) is directly proportional to the relative range between the current distance of the human from the robot and the neutral distance, as shown in Figure 9.

The maximum translational velocity is defined as 0.6 m/s. Similarly, the rotational velocity (rv) is directly proportional to the angular difference between the predefined neutral direction and the current direction of the targeted person, and the maximum angular velocity is defined as 0.9 rad/s. The rest of the parameters

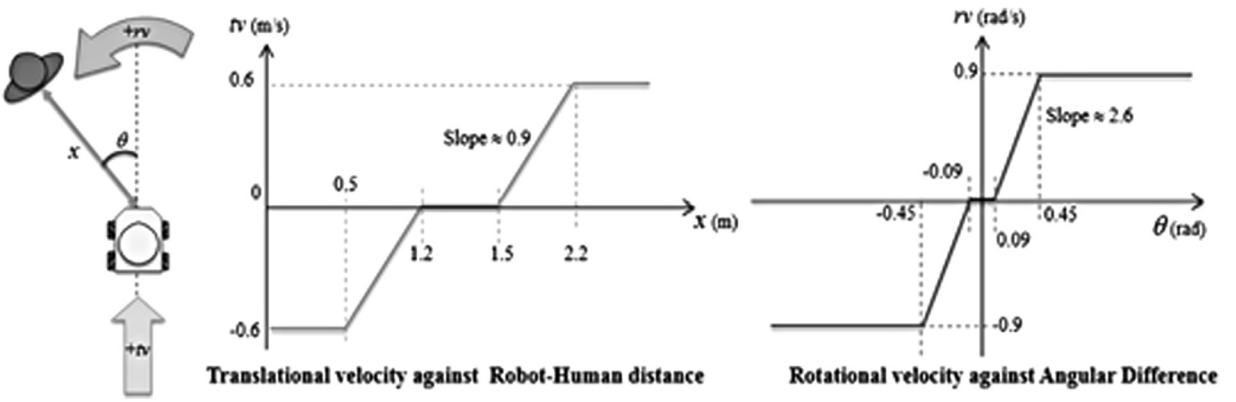

Figure 9. *Velocity profiles for following behavior.*

were tuned empirically in order to obtain the velocity profiles for *tv* and *rv*, as shown in Equation 6 and Equation 7, respectively. In this behavior, the strafe velocity (*sv*) is not used to generate the motion for the person-following behavior.

$$tv = 0.6\,f(x) \text{ where } f(x) = \begin{cases} -1 & \text{if } x < 0.5 \\ \frac{1.2-x}{0.7} & \text{if } 0.5 \le x < 1.2 \\ 0 & \text{if } 1.2 \le x \le 1.5 \\ \frac{x-1.5}{0.7} & \text{if } 1.5 < x \le 2.2 \\ 1 & \text{if } x > 2.2 \end{cases} \tag{6}$$

$rv = 0.9f(\theta)$ where

$$f(\theta) = \begin{cases} -1 & \text{if } \theta < -0.45 \\ \frac{-0.09-\theta}{-0.36} & \text{if } -0.45 \le \theta < -0.09 \\ 0 & \text{if } -0.09 \le \theta \le 0.09 \\ \frac{\theta-0.09}{0.36} & \text{if } 0.09 < \theta \le 0.45 \\ 1 & \text{if } \theta > 0.45 \end{cases} \tag{7}$$

4.2.2 Side-by-Side Accompaniment. In the accompaniment behavior, the Kinect sensor was mounted at the side of the robot such that the Z_0 axis of the sensor pointed to the left-hand-side direction of the robot. In this manner, the sensor was able to track the interactant while the interactant walked beside the robotic avatar on its left-hand side. The setup is as shown in Figure 10(a), where the relative position between the robot and the human is described. Three quantities, including the robot–human distance, the rotation angle, and the offset were used in the computations of velocity commands for the accompaniment behavior. The robot–human distance and offset distance can be obtained directly from the Kinect SDK.

Figure 10(b) illustrates the coordination system of the interactant's human frame with respect to the Kinect frame. The Kinect frame is X_0-Υ_0-Z_0, and the human frame is X_1-Υ_1-Z_1. The rotation matrix from X_0-Υ_0-Z_0 to X_1-Υ_1-Z_1 was obtained with the Microsoft Kinect SDK, and it is of the following format:

$$R_0^1 = \begin{bmatrix} M_{11} & M_{12} & M_{13} & 0 \\ M_{21} & M_{22} & M_{23} & 0 \\ M_{31} & M_{32} & M_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

The 3-by-1 vectors $\begin{pmatrix} M_{11} \\ M_{21} \\ M_{31} \end{pmatrix}$, $\begin{pmatrix} M_{12} \\ M_{22} \\ M_{32} \end{pmatrix}$, and $\begin{pmatrix} M_{13} \\ M_{23} \\ M_{33} \end{pmatrix}$ are the unit vectors of the axes X_1, Υ_1, and Z_1, respectively, and they represent the orientation of X_1, Υ_1, and Z_1 with respect to the Kinect frame. In the accompaniment mode, the direction at which the interactant is facing is that of Z_1. Therefore, the unit vector of the Z_1 axis is used to calculate the rotation angle that the robot must rotate so that it faces the same direction as the interactant it is following, as shown in Equation 9.

$$\theta = \tan^{-1}\left(\frac{M_{33}}{-M_{13}}\right) \tag{9}$$

The velocity commands that produce the motion of the robot during the side-by-side accompaniment include translational velocity (*tv*), strafe velocity (*sv*), and rotational velocity (*rv*). The offset distance (*y*) will

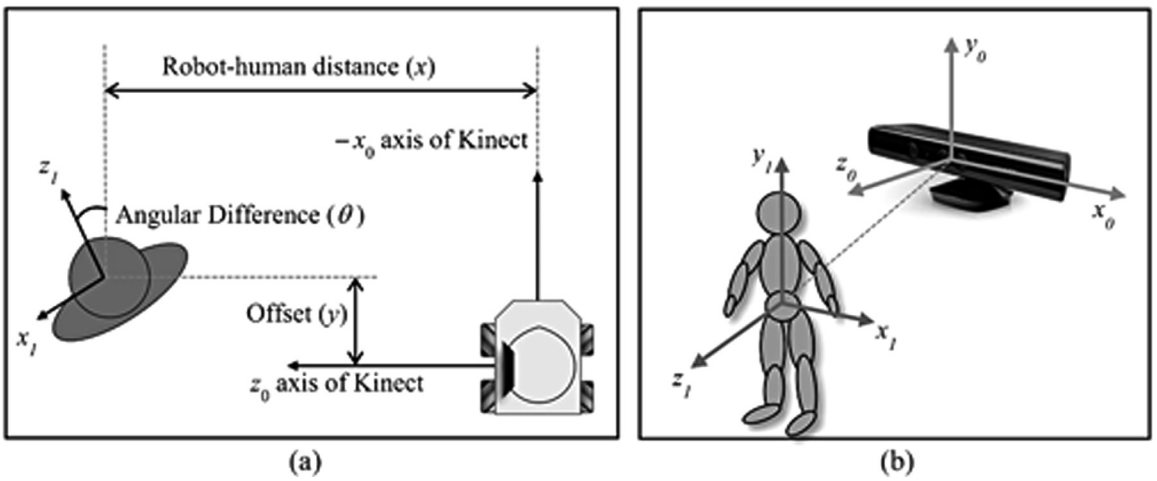

Figure 10. *(a) Relative robot–human positions in side-by-side accompaniment; (b) human frame and Kinect frame.*

determine the translational velocity and the angular difference (θ) will influence the rotational velocity, while the strafe velocity is dependent on the robot–human distance. The *tv*-*y*, *rv*-θ, and *sv*-*x* relationships are directly proportional and they are depicted in Equations 10, 11, and 12, respectively.

$tv = 0.6\,f(y)$ where

$$f(y) = \begin{cases} -1 & \text{if } y < -0.55 \\ \frac{0.15+y}{0.4} & \text{if } -0.55 \le y < -0.15 \\ 0 & \text{if } -0.15 \le y \le 0.15 \\ \frac{y-0.15}{0.4} & \text{if } 0.15 < y \le 0.55 \\ 1 & \text{if } y > 0.55 \end{cases} \quad (10)$$

$rv = 0.2\,f(\theta)$ where

$$f(\theta) = \begin{cases} -1 & \text{if } \theta < -0.81 \\ \frac{0.52+\theta}{0.29} & \text{if } -0.81 \le \theta < -0.52 \\ 0 & \text{if } -0.52 \le \theta \le 0.52 \\ \frac{\theta-0.52}{0.29} & \text{if } 0.52 < \theta \le 0.81 \\ 1 & \text{if } \theta > 0.81 \end{cases} \quad (11)$$

$sv = 0.6\,f(x)$ where

$$f(x) = \begin{cases} -1 & \text{if } x < 1.2 \\ \frac{1.6-x}{0.4} & \text{if } 1.2 \le x < 1.6 \\ 0 & \text{if } 1.6 \le x \le 1.8 \\ \frac{x-1.8}{0.4} & \text{if } 1.8 < x \le 2.2 \\ 1 & \text{if } x > 2.2 \end{cases} \quad (12)$$

4.2.3 Guiding Behavior. In the guiding behavior, the robotic avatar moved in front of the interactant and its task was to escort the interactant around. Therefore, unlike the following and the accompaniment behaviors, the robot in this behavior moved to a designated goal position, which is independent of the interactant's movements. The robotic avatar planned a path to the goal position and moved toward the destination at a translational velocity of V_T and an angular velocity of V_R.

Although the robot knew where to go, it tracked the interactant so that it could adjust its velocity according to the interactant's movement. The robot still tried to maintain a predefined 1.5 m distance from the interactant. If the interactant lagged behind and the distance between the interactant and robot was greater than 1.5 m, the robot slowed down to accommodate the interactant. On the other hand, if the interactant increased his or her walking speed, the robot moved more quickly toward the goal position. In this manner, the final velocity commands, *tv* and *rv*, are proportional to the distance of the person from the robot (x), as shown in Equation 13.

$$\begin{aligned} tv &= V_T\,f(x) \\ rv &= V_R\,f(x) \end{aligned} \quad \text{where } f(x) = \begin{cases} \frac{1.2}{x} & \text{if } x < 1.2 \\ 1 & \text{if } 1.2 \le x \le 1.5 \\ \frac{1.5}{x} & \text{if } 1.5 < x < 2.0 \\ 0 & \text{if } x \ge 2.0 \end{cases} \quad (13)$$

4.3 Obstacle Avoidance

Currently, the obstacle avoidance capability has only been implemented for the following and guiding behaviors. In the accompaniment behavior, the robot would temporarily switch to the following behavior until the robot ascertains that adequate space is available. It would then switch to executing the accompaniment behavior. The reactive obstacle avoidance capability has been implemented using the concept of the virtual potential field (VPF; Koren & Borenstein, 1991). The virtual repel force exerted by the obstacle is inversely proportional to the distance (or distance raised to a certain power) between the obstacle and the moving robot. That is, if the robot comes closer to the obstacle, the virtual repel force on it from the obstacle will be larger. The robot would then move away from the obstacle with a higher velocity.

While the Kinect sensor is used for tracking the interactant, it is also used to observe obstacles between the robot and the interactant. The Kinect sensor is able to obtain the 3D position of all points on the depth image. The system would first calculate the current distance from the robot to the interactant. For each pixel on the Kinect sensor's depth image that corresponds to a distance less than that between the robot and the interactant, the system would consider that as belonging to an obstacle. A virtual repel force is then exerted on the robot by that point. After all obstacle points have been identified, the virtual repel forces contributed by these points are then summed up. Finally, the summation of virtual repel forces is divided by the total number of obstacle points, resulting in the average virtual repel force, which would then be used to calculate the robot's velocity in order to avoid obstacles. The virtual repel force, F_{repel}, from each obstacle point is given by Equation 14, where K is a constant and D is the distance from the point to the robot.

$$F_{\text{repel}} = \frac{K}{D^3} \quad (K = 0.6) \tag{14}$$

For a small value of D, the resulting virtual repel force will be larger. In order for the robot to efficiently avoid obstacles during the front-following and the back-following modes, the robot's strafe velocity (sv) was calculated from the average virtual repel force and then transmitted to MAVEN-II. The translational velocity (tv) and rotational velocity (rv) were not affected by the existence of obstacles; they were determined only by the position of the followed person relative to the robot. When tv, sv, and rv are combined, the resulting motion of the robot enables obstacle avoidance. At the same time, the robot would also be able to maintain a desirable distance and orientation toward the interactant. The virtual repel force is proportional to the incremental amount of transverse speed sv, as shown in Equation 15, where A is a predefined coefficient, and $\triangle sv$ represents the incremental sv value after each update of image and depth frame from the Kinect sensor.

$$\triangle sv = A \times F_{\text{repel}} \tag{15}$$

The resulting transverse robot speed can be calculated using Equation 16, where A is set to 1 for simplicity and n is the current number of Kinect data frames, counting from the moment when an obstacle has been detected (when $n = 0$).

$$sv = \triangle sv \times n = A \times F_{\text{repel}} \times n \tag{16}$$

The sv value will gradually increase with the increasing value of n, so that the obstacle avoidance movement will not be too abrupt. Since the frame update rate of Kinect is fast (minimum 9 Hz), the sv value can increase from zero up to its maximum speed (0.3 m/s) within 1 s. This would allow the robot to quickly avoid nearby obstacles.

5 Experiment Evaluation

The autonomous multimodal person-following system has been implemented and the robustness of the following behavior as well as the accompaniment behavior has been validated in Pang, Seet, and Yao (2013). In this paper, the study seeks to address the objectives of equipping the robotic avatar with autonomy in navigation, as outlined earlier in Section 2. Experiments have been carried out to evaluate the following hypotheses.

Hypothesis 1: The addition of autonomous navigation behaviors to robotic avatar helps to reduce the inhabitor cognitive workload.
Hypothesis 2: The addition of autonomous navigation behaviors to a robotic avatar helps to ensure safety during the navigation of the robotic avatar in a remote environment.

5.1 Experiment Design

5.1.1 Participants. A total of five volunteers (5 males; mean age = 30.2 years, $SD = 4.9$) were sourced to become unpaid participants for the experiment. Participants would assume the role of the inhabitor in the telepresence system. The number of participants was small due to the time necessary for participant training to ensure adequate fluency in deploying the robotic avatar prior to conducting the actual test sessions.

5.1.2 Familiarization Session. The experiments were conducted in an indoor laboratory environment. A familiarization session was conducted to acquaint the participants with the laboratory and the inhabitor station. The inhabitor station was placed in an enclosed room and it was composed of a laptop that was connected to the internet, as well as a gamepad that was used to control the robot manually. The participant would control the robotic avatar from a webpage loaded in a web browser. Live video and audio feeds from the robotic avatar's webcam were displayed on the webpage. All participants were tested individually. Each participant completed a 2-min warm-up drive while the research assistant was in the room. The research assistant guided the participants regarding the use of the interface during the warm-up drive and answered any questions about the tasks.

The robotic avatar and participants were located in separate rooms that were approximately 20 m apart. During each run, the participant was left alone. The research assistant adopted the role of the interactant. As such, the research assistant was able to see and hear the participant via the robotic avatar.

5.1.3 Tests. There were two tests for each participant. The first served to evaluate the participants' performance while operating the robotic avatar manually with a gamepad. The second test served to evaluate their performance while the robot operated in the autonomous configuration.

For both tests, the participant was instructed to control the robotic avatar and to follow the interactant. The interactant would move along a predefined path. Markers were affixed on the floor to ensure that the interactant followed the predefined path, as shown in Figure 11. The interactant tried to have a conversation with the participant during the experiment, simulating a walk-and-talk scenario using a robotic avatar. The NASA-Task Load Index (TLX; Hart & Staveland, 1988) was used to provide understanding of participant workload. It is a questionnaire that was administered at the end of the test sessions and rated workload according to six dimensions: mental workload, physical demand, temporal demand, the perceived level of task performance by the participant, effort, and frustration. The participant would rate each of these dimensions along a scale with 10 divisions. Each division represents a 10-point increment, effectively allowing the participant to rate the dimension with a score ranging from 0 to 100. The participant's overall workload was derived with the mean score from all six dimensions (Young & Stanton, 2004).

The performance measures in this experiment also included the time taken to complete the task of following an interactant as well as the total number of collisions that occurred over the course of performing the task.

6 Results

The results of the experiment are presented in Table 1.

The mean workload and the mean time taken to complete the task, as well as the total number of collisions, were found to be lower when participants operated the robotic avatar in the autonomous mode.

The mean workload was 45.7% higher when the robotic avatar was operated to follow an interactant in

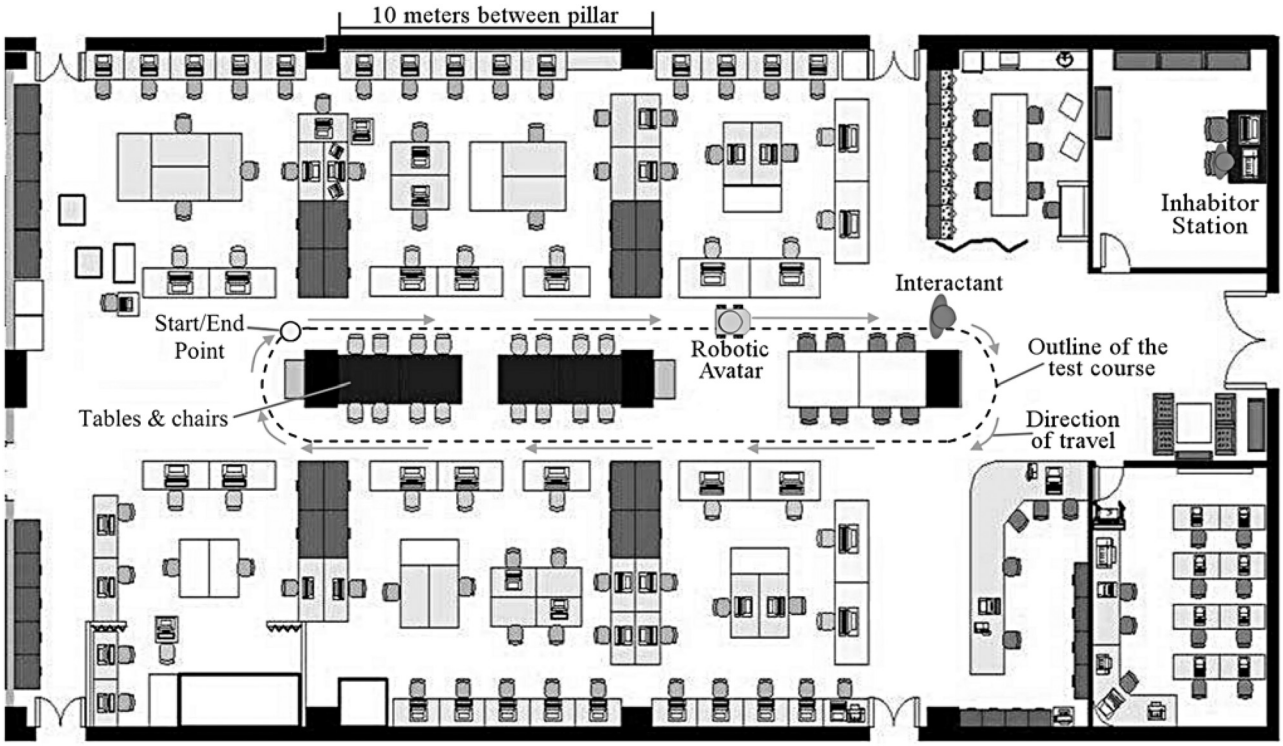

Figure 11. *Schematic layout of the test environment.*

Table 1. *Experimental Results*

	Manual			Autonomous		
	Mean NASA-TLX	Mean time taken (s)	Number of collisions	Mean NASA-TLX	Mean time taken (s)	Number of collisions
Participant 1	31.67	120.09	2	36.67	120	0
Participant 2	53.33	155.003	0	18.33	100.07	0
Participant 3	50	266.046	1	18.33	144.053	0
Participant 4	26.67	136.087	0	48.33	111.007	0
Participant 5	35	156.036	1	13.33	97.013	0
Mean	39.33	166.652	0.8	27	114.429	0
SD	11.7	57.5104		14.88	18.9065	

the manual mode, compared to when performing the same task in the autonomous configuration. This result provides considerable evidence in support of Hypothesis 1, which states that the cognitive workload of the inhabitor is lower when the robotic avatar is in the autonomous configuration. All the participants except for Participant 4 experienced a lower level of stress when operating the robot in the autonomous mode. That participant commented that when the robotic avatar was in autonomous mode, it traveled at a higher speed and caused a lot of noise. The noise made it difficult for the participant to hear the interactant and resulted in an

inability to have a proper conversation via the robotic avatar.

All participants took less time to complete the task when they were operating the robotic avatar in the autonomous configuration. The mean time taken to complete the task is 31.3% lower when the robot is in autonomous mode. Participant 3 took a significantly longer time to complete the test because of his familiarity with teleoperating another nonholonomic mobile robotic platform. The participant explained that this familiarity may have interferred with his performance in this experiment.

Lastly, when the robotic avatar was controlled in the manual mode, there were a total of four collisions. On the other hand, no collisions occurred when the robotic avatar was operated in the autonomous configuration. This supports Hypothesis 2, which states that the addition of autonomous navigation behaviors to the robotic avatar helps to increase safety during the navigation of the robotic avatar in a remote environment.

The results have helped to reinforce the belief that the addition of higher-level navigational behaviors can reduce cognitive workload while also enhancing safety during the navigation of the robotic avatar in a remote environment.

7 Conclusion

This paper reports on the development work on implementing a robot-mediated telepresence application, called MAVEN. MAVEN exists in two versions: a flat screen robotic avatar and humanoid robotic avatar.

A framework has been proposed to highlight the need for higher levels of navigational autonomy. It has been proposed that high levels of navigational autonomy would allow for the inhabitor to also provide inputs for head and arm gesticulation. This is because the inhabitor's arms would be freed from the task of manipulating input devices for expressing commands for robot navigation. Furthermore, the inhabitor can focus on looking at the camera for communication with interactants rather than being engrossed in monitoring the robot navigation.

This paper has discussed the development of these higher-level navigational behaviors. These behaviors are following, accompaniment, and guiding. Each of these navigation behaviors assists in navigating the robot during a certain type of interaction that is typical between two human individuals.

The experiment demonstrated that the following behavior has been implemented robustly, with a small chance of losing track of the interactant. The experiment results have also shown that the high-level navigational behaviors help to reduce workload. Furthermore, a smaller number of collisions was recorded when the robot was performing an autonomous behavior.

Future work would include designing and implementing more human-like navigational behaviors within the framework. The autonomous multimodal person-following system can be modified to use one Kinect sensor instead of three sensors. The behaviors can also be enhanced by taking into account more parameters for velocity profiling.

Acknowledgments

This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Program.

References

Anybots. (2010). *Anybots virtual presence systems. It's you, anywhere...* Retrieved from https://www.anybots.com/

Arras, K. O., Mozos, O., & Burgard, W. (2007). Using boosted features for the detection of people in 2D range data. In *Proceedings of the IEEE/RSJ International Conference on Robotics and Automation, ICRA*, 3402–3407.

Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. Sebastopol, CA: O'Reilly Media.

Burgard, W., Cremers, A. B., Fox, D., Hahnel, D., Lakemeyer, G., Schulz, D., ... Thrun, S. (1998). The interactive

museum tour-guide robot. *Proceedings of the 15th National Conference on Artificial Intelligence*, 11–18.

Chen, Z. C., & Birchfield, S. T. (2007). Person following with a mobile robot using binocular feature-based tracking. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 815–820.

Cosgun, A., Florencio, D. A., & Christensen, H. I. (2013). Autonomous person following for telepresence robots. *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA*, 4335–4342.

Digital Video Enterprises. (2010). *DVE Immersion Room*. Retrieved from http://dvetelepresence.com/room/home.htm

Doisy, G., Jevtic, A., Lucet, E., & Edan, Y. (2012). Adaptive person-following algorithm based on depth images and mapping. *Proceedings of the Workshop on Robot Motion Planning: Online, Reactive, and in Real-time, IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012.*

Ellison, L. M., Pinto, P. A., Kim, F., Ong, A. M., Patriciu, A., Stoianovici, D., ⋯ Kavoussi, L. R. (2004). Telerounding and patient satisfaction after surgery. *Journal of the American College of Surgeons, 199*(4), 523–530.

Fod, A., Howard, A., & Mataric, M. J. (2002). A laser-based people tracker. *Proceedings of the IEEE/RSJ International Conference on Robotics and Automation, ICRA*, 3024–3029.

Gockley, R., Forlizzi, J., & Simmons, R. (2007). Natural person-following behavior for social robots. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, HRI* 17–24.

Hall, E. T. (1990). *The hidden dimension*. New York: Anchor Books.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *International Journal of Advances in Psychology*, Vol. 52, 139–183.

In Touch Technologies. (2011). *RP endpoint devices.* Retrieved from http://www.intouchhealth.com/

Kirby, R., Simmons, R., & Forlizzi, J. (2009). Companion: A constraint-optimizing method for person-acceptable navigation. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 607–612.

Koren, Y., & Borenstein, J. (1991). Potential field methods and their inherent limitations for mobile robot navigation. *Proceedings of the IEEE/RSJ International Conference on Robotics and Automation, ICRA*, Vol 2, 1398–1404.

Kwon, H., Yoon, Y., Park, J. B., & Kak, A. C. (2005). Person tracking with a mobile robot using two uncalibrated independently moving cameras. *Proceedings of the IEEE/RSJ International Conference on Robotics and Automation, ICRA*, 2877–2883.

Lee, M. K., & Takayama, L. (2011). Now, I have a body: Uses and social norms for mobile remote presence in the workplace. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI*, 33–42.

Lincoln, P., Welch, G., Nashel, A., Ilie, A., & Fuchs, H. (2009). Animatronic shader lamps avatars. *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality, ISMAR*, 27–33.

Loper, M. M., Koenig, N. P., Chernova, S. H., Jones, C. V., & Jenkins, O. C. (2009). Mobile human-robot teaming with environmental tolerance. *Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction, HRI*, 157–164.

Luber, M., Spinello, L., & Arras, K. O. (2011). People tracking in RGB-D data with on-line boosted target models. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 3844–3849.

Minsky, M. (1980). Telepresence. *Omni, 2*(9), 45–52.

Montemerlo, M., Pineau, J., Roy, N., Thrun, S., & Verma, V. (2002). Experiences with a mobile robotic guide for the elderly. *Proceedings of the 18th National Conference on Artificial Intelligence*, 587–592.

Morales, Y., Satake, S., Huq, R., Glas, D., Kanda, T., & Hagita, N. (2012). How do people walk side-by-side? Using a computational model of human behavior for a social robot. *Proceedings of the 7th ACM/IEEE International Conference on Human–Robot Interaction, HRI*, 301–308.

Ohya, A., & Munekata, T. (2002). Intelligent escort robot moving together with human-interaction in accompanying behavior. *Proceedings of FIRA Robot World Congress*, 31–35.

Pacchierotti, E., Christensen, H. I., & Jensfelt, P. (2006). Design of an office-guide robot for social interaction studies. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 4965–4970.

Pang, W. C., Burhan, B., & Seet, G. (2012). Design considerations of a robotic head for telepresence applications. In *Intelligent robotics and applications. Lecture notes in computer science*, Vol. 7508, pp. 131–140. Berlin: Springer-Verlag.

Pang, W. C., Seet, G., & Yao, X. (2013). A multimodal person-following system for telepresence applications. *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology, VRST*, 157–164.

Prassler, E., Bank, D., Kluge, B., & Hagele, M. (2002). Key technologies in robot assistants: Motion coordination between a human and a mobile robot. *Transaction on Control, Automation, and Systems Engineering, 4*(1), 56–61.

Seet, G., Pang, W. C., & Burhan, B. (2012). Towards the realization of MAVEN: Mobile robotic avatar. *Proceedings of the 25th International Conference on Computer Animation and Social Agents.*

Seet, G., Pang, W. C., Burhan, B., Chen, I.-M., Viatcheslav, I., William, G., & Wong, C. Y. (2012). A design for a mobile robotic avatar: Modular framework. *Proceedings of the 3DTV Conference: The true vision; Capture, transmission and display of 3D video, 3DTV-CON.*

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments, 1*(1), 120–126.

Szigeti, T., McMenamy, K., Saville, R., & Glowacki, A. (2009). *Cisco telepresence fundamentals.* Indianapolis, IN: Cisco Press.

Thacker, P. D. (2005). Physician-robot makes the rounds. *Journal of the American Medical Association, 293*(2), 150.

Topp, E. A., & Christensen, H. I. (2005). Tracking for following and passing persons. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2321–2327.

Tsui, K. M., Desai, M., Yanco, H. A., & Uhlik, C. (2011). Exploring use cases for telepresence robots. *Proceedings of the 6th ACM/IEEE International Conference on Human–Robot Interaction, HRI*, 11–18.

Turletti, T., & Huitema, C. (1996). Videoconferencing on the internet. *IEEE/ACM Transactions on Networking, 4*(3), 340–351.

Venolia, G., Tang, J., Cervantes, R., Bly, S., Robertson, G., Lee, B., & Inkpen, K. (2010). Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI*, 1049–1058.

Walker, G. R., & Sheppard, P. (1997). Telepresence: The future of telephony. *BT Technology Journal, 15*(4), 11–18.

Willow Garage. (2011). *Texai remote presence system.* Retrieved from http://www.willowgarage.com/pages/texai/overview

Young, M. S., & Stanton, N. A. (2004). Taking the load off: Investigations of how adaptive cruise control affects mental workload. *Ergonomics, 47*(9), 1014–1035.