Autonomous Performance of Multistep Activities with a Wheelchair Mounted Robotic Manipulator Using Body Dependent Positioning

Hairong Jiang, *Student Member, IEEE,* Ting Zhang, *Student Member, IEEE,* Juan P. Wachs, *Member, IEEE,* and Bradley S. Duerstock*, *Member, IEEE*

***Abstract*—In this paper, an autonomous vision-based system was developed to control a wheelchair mounted robotic manipulator (WMRM). Two 3D cameras were applied for object and body part recognition (face and hands) of the wheelchair user. Two human robot interface modalities were used to control the WMRM: voice and gesture recognition. Daily objects were automatically recognized by employing a two-step process: 1) using Histogram of Oriented Gradients (HOG) algorithm to extract the feature vector for each detected object; 2) applying nonlinear support vector machine (SVM) algorithm to train the model and classify the objects. Four simulated tasks for daily objects delivery and retrieval were designed to test the validity of the proposed system. The results demonstrated that the automatic control requires significantly fewer time than the predefined control for phone calling and photography tasks (*$P = 0.015$, $P = 0.035$*), respectively. The gesture modality outperforms the voice control for the drinking and phone calling tasks (*$P = 0.016$, $P = 0.015$*), respectively.**

I. Introduction

The advancement of assistive robotics facilitates the development of wheelchair mounted robotic manipulators (WMRMs) for people with disabilities (PWDs). These WMRMs worked in close proximity to PWDs to assist with Activities of Daily Living (ADL), such as dressing, feeding, and objects retrieval and delivery. WMRMs improve the accessibility of surroundings for PWDs and enhance their independence [1].

Previous research has shown that a WMRM system is beneficial to individuals with mobility impairments, such as spinal cord injury (SCI) [2] and Cerebral Palsy [3]. An intelligent assistive robotic manipulator system named UCF-MANUS was developed by Kim et al. [4] for users with a wide range of disabilities. Essential to this system and other WMRMs is to integrate computer vision to recognize daily objects. For example, Fence et al. [5] applied a monocular camera for object recognition using scale invariant feature transform (SIFT) to control a 7-degree of freedom (DoF) robotic arm. The parts of the body of the operator were also recognized to assist in automating daily tasks. Tanaka et al. [6] developed an assistive WMRM to grasp a cup and bring it to the user's mouth with the help of two cameras (one is used to recognize the objects and the other is used to recognize the user's face) in the robotic arm's hand.

Recently, the availability of commercial WMRMs has increased. For example, JACO robotic manipulator produced by Kinova® and Cyton Gamma 1500 [7] developed by ROBAI® (a lightweight 7-DoF robotic arm) are designed to be mounted on the wheelchair and help users with upper limb impairments with instrumental and basic ADLs [8].

Previous studies on human robot interfaces (HRIs) for robotic manipulator control were based on different input modalities. For instance, Pathirage *et al*., [9] developed a vision-based Brain Computer Interface (BCI) to grasp objects using a WMRM. The patients with tetraplegia were trained to voluntarily modulate electroencephalogram (EEG) signals to send commands to a WMRM. Three modalities were adopted in the WMRM system presented by Kim et al. including joystick, touchscreen, and BCI [4]. Other control modalities for the WMRM systems consist of speech recognition [10], head movement and facial expression [11], hand gestures [12], EEG signals [13], and a 3-D controller [14].

Our previous work consists of designing a gesture recognition-based interface for quadriplegic individuals due to SCI [12] and developing a prototype vision-based WMRM system (with manual and semi-automatic control mode) combining hand gestural control and automatic user face and object detection for quickly retrieving everyday objects for use [15]. The drawback of the previous developed system was that it required the user to manually control the robotic arm to perform fine movements when grasping an object.

In this paper, we extend the functionality and robustness of the object recognition algorithm, the HRI modalities we test, and the robotic control policy used. Integrated computer vision algorithms are applied to detect, recognition, and grasp objects automatically. The human body parts (face and hands) are tracked to facilitate objects positioning. Additionally, approximation signals from the smartphone are used to provide feedback for the users' safety and tasks' efficiency. Moreover, the system was tested with more complex, multistep tasks to simulate real-world needs.

II. System Architecture

The architecture of this prototype system is illustrated in Fig. 1. The computer vision-based WMRM system includes five modules: (A) user interface with gesture and speech control, (B) automatic object recognition, (C) human body part recognition, (D) object sensors, and (E) the robotic arm control module. Four multistep tasks were designed to test this system: drinking, phone calling, taking a self-portrait or 'selfie' photograph, and typical picture taking.

* This research is supported by the State of Indiana to the Center for Paralysis Research.

H. Jiang, T. Zhang, and J. P. Wachs are with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: jiang115@purdue.edu, zhan1013@purdue.edu, jpwachs@purdue.edu).

B. S. Duerstock is with the Weldon School of Biomedical Engineering and School of Industrial Engineering, Purdue University, West Lafayette, IN 47906 USA (765-496-2364; e-mail: bsd@purdue.edu).

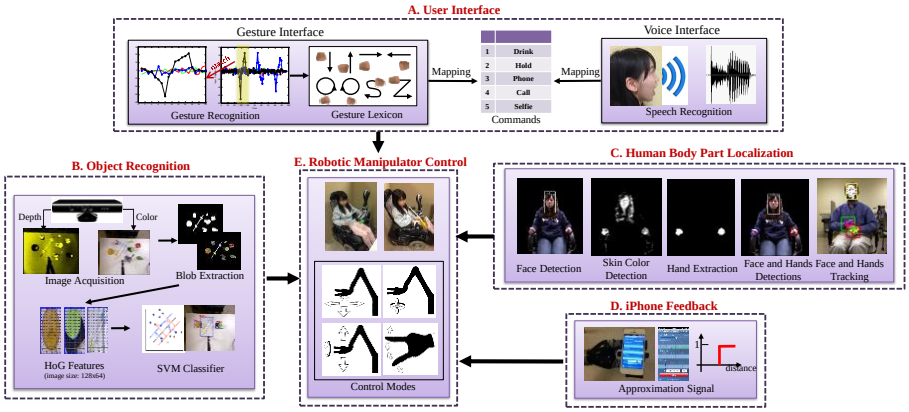


Figure 1. System Architecture

A. *User Interface Module*

The target population of this WMRM system is users with upper extremity mobility impairments who lack fine motor skills. The commonly used input modalities, such as keyboard, joystick, and touchscreen, require the users to make fine motor movements and physically contact these interfaces. This can be difficult for quadriplegics with severe disabilities, thus they were not selected [14]. Two control modalities were adopted: 1) gesture recognition based interface and 2) speech recognition based interface.

The gesture recognition based interface consists of three parts: foreground segmentation, hand detection and tracking, and trajectory recognition. A detailed description can be referred to [12]. In the foreground segmentation stage, two steps were applied to segment the human body and its connected components as the foreground. The first step consists of thresholding the depth image (acquired from a Kinect camera) using two thresholds (T_{DH} and T_{DL}). T_{DH} and T_{DL} are the high and low thresholds for depth value, respectively [12]. A binary mask image is then generated with each pixel's depth between T_{DH} and T_{DL}. The largest blob in the binary mask is extracted as the foreground and all the remaining blobs are removed. In the hand detection stage, a face detector and a skin color detector is applied to detect the face and both hands from the foreground. A particle filter framework incorporating motion and spatial information is applied to track the hands (Fig. 1-C). In the recognition stage, the hand trajectories are recognized by the CONDENSATION algorithm. The state S at time t was extended to recognition two hands' trajectories (Eq. 1)

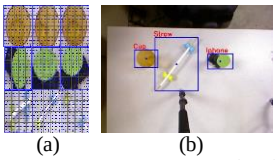
$$S_t = (\mu, \phi^i, \alpha^i, \rho^i) = (\mu, \phi^{right}, \phi^{left}, \alpha^{right}, \alpha^{left}, \rho^{right}, \rho^{left}) \quad (1)$$

where, μ is the index of the motion models, ϕ is the current phase in the model, α is an amplitude scaling factor, ρ is a time dimension scaling factor, i equals to right hand, or left hand. An eight-gesture lexicon was adopted for this gesture recognition based interface to control the WMRM.

For speech recognition, the CMU sphinx was adopted. It is an open source toolkit for speech recognition and has been widely used [16]. The system segments input voice signals into words and then compare them with the pre-trained model. When a key word is recognized, a corresponding command is sent to control the robotic arm. Although speech recognition based interface may not be ideal with a noisy environment, it does not require any movement from the hands and can work well in the indoor quiet environment. Thus, it is selected as a complement to the gesture-based interface to satisfy the needs of users with different levels of mobility impairments.

B. *Object Recognition Module*

In this work, the performance of the object recognition module is improved by incorporating the depth information of the Kinect camera and a combination of machine learning algorithms. An example of the color and depth information captured by a Kinect camera is shown as in Fig. 1-B (upper left). A depth threshold is applied to segment each object as a blob. The region of interest (ROI) consisting of the detected object is separated from the original image and resized to 128x64. The histogram of oriented gradients (HOG) features (edge gradients and orientations in Eq. 2 & 3) are extracted from each resized image (Fig. 2a). x and y are pixel values, m and θ are the magnitude and orientation of the gradient, respectively. The features are trained and classified using a nonlinear Support Vector Machine (SVM) algorithm [17] (Fig. 2b).



(a) (b)

Figure 2. (a) Feature Extraction; (b) SVM classifier for object recognition

$$m(x,y) = \sqrt{dx(x,y)^2 + dy(x,y)^2} \quad (2)$$

$$\theta(x,y) = \begin{cases} \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) - \pi & if\ dx(x,y) < 0\ and\ dy(x,y) < 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) + \pi & if\ dx(x,y) < 0\ and\ dy(x,y) > 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) & otherwise \end{cases} \quad (3)$$

C. *Human Body Part Localization Module*

For different users or different postures of the same user, the destination for objects delivery could be different. An anthropometric relationship is measured to properly position objects to the user. The position of the face and hands are tracked and applied to automatically position a cup with a straw, mobile phone and other daily living objects to the user. For example, the phone is positioned to the position near the

hand, so that the user can dial and then placed near the ear to have a private conversation.

D. iPhone Feedback Module

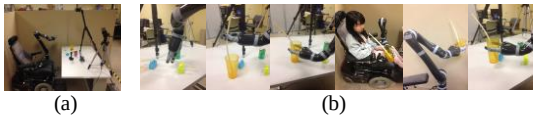
This system was integrated with three of the selected experimental tasks (phone calling, selfie, and picture taking task) using the Apple® iPhone™. To provide sensory input for the calling task, we used the iPhone application named Sensor Streamer [18]. The signal of the built in Proximity sensor is sent as feedback to control the WMRM. When the iPhone is within a certain distance (obtained from the Proximity sensor) from the users' ear, the WMRM is commanded to stop to begin a phone conversation.

E. Robotic Manipulator Control Module

A 6-DOF commercial robotic arm produced by Kinova Robotics is adopted and mounted on the left side of an electric wheelchair to enable interaction and manipulation. The JACO API was programmed under C# environment as a wrapper. The speech and gesture recognition, and body part tracking results were then sent as commands to control the robotic arm. Two control modes were tested: autonomous (the objects and human body parts were automatically detected and localized) and predefined position control (the objects' and participants' locations were predefined).

III. Experimental Results

Four multistep tasks were designed to test the proposed system: 1) drinking with a straw; 2) making a private phone call, 3) taking self-portrait or 'selfie' photos; and 4) taking photos of the surroundings. The settings of the experiments are illustrated in Fig. 4. Two Kinect cameras were used: one facing the objects of interest on a table and another facing the user for gesture recognition and human body part localization (Fig. 3a). The four tasks were assessed by a participant during five trials.



(a) (b)

Figure 3. Experimental setting and procedure. (a) Experiment setting; (b) Operations of the drinking task.

A. Task 1: Getting the users to drink with a straw

The drinking task consists of six operations: 1) picking up a long straw from a table, 2) putting the straw into a cup, 3) picking up the cup, 4) delivering the cup to the front of the participant for drinking, 5) holding the cup to the side of the participant at rest, and 6) placing the cup back on the table when finished (Fig. 3b). Task completion time and the error rate for all the six steps are recorded. The task completion time begins with operation 1) and ends when operation 6) finished. The error rate represents the ratio between the number of failed operations and the number of total operations (6). Fig. 4 shows the results that the average task completion time over five trials of gesture modality is significantly less than voice modality using a predefined positioning approach ($P = 0.016$). A two-way ANOVA shows that there was no statistically significant interaction between gesture and voice input modalities and robot positioning (predefined/autonomous) for task completion time ($P = 0.205$). Gesture recognition showed much greater error rate than voice control. On average there was not much difference in accuracy between autonomous and predefined positioning. For predefined positioning using voice control, longer completion time resulted in the lowest error rate. A time/error correlation coefficient was calculated to determine whether lower accuracy was related to longer performance time. A negative coefficient value of -0.43 indicates only a mild inverse relationship between these two factors.

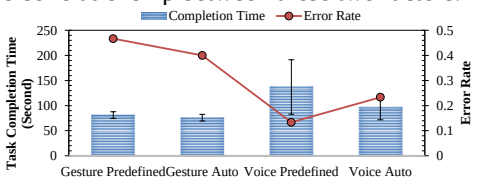


Figure 4. Task completion time and accuracy for the drinking task

B. Task 2: Making a private phone call

Making a phone call is a basic ADL. However, holding the phone to one's ear is difficult for individuals with upper extremity mobility impairments. Either they must wear a headset or use the speakerphone function, which lacks privacy. This task enables operators to pick up a mobile smartphone from the table with the WMRM, putting it in dialing position, and then placing it to the user's ear. The completion time recorded in this task was the time duration from when participants finished dialing to when the phone stopped beside participants' left ear (Fig. 5).

Participants were asked to exchange a random 6-digit code of both numbers and letters with the experimenter to verify proper hearing through the phone. The number of wrong digits received are used to compute errors. From t-test results, the task completion time of the automatic positioning requires significantly less time than predefined positioning ($P=0.015$) (Fig. 5). In Fig. 5, the predefined positioning with gesture control shows longer completion time with lower error rate compared to automated positioning using either gesture and voice interface. The time/error correlation coefficient of 0.011 indicates the higher accuracy is not likely a direct consequence of longer completion time.

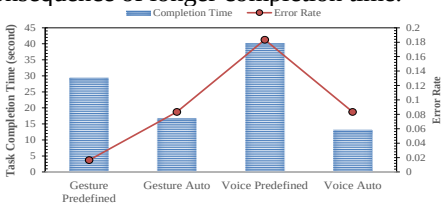


Figure 5. Task completion time and accuracy for phone calling task

C. Task 3: Taking pictures of the surroundings

The picture taking task consists of picking up the smartphone within the participant's reach and allowing them to take a picture using camera feature. In Fig. 6, the average completion time indicates the time duration to adjust the phone's position to take a satisfactorily framed photograph of objects on the wall in front of the participant. The results of the Tukey's post-hoc test indicate that automatic gesture control requires significantly more time to complete than the other control methods ($P = 0.049, 0.032, 0.036$). Correct positioning of the phone depended on the quality of the photo taken by through the camera lens rather than its orientation to the participant's body.

D. *Task 4: Taking selfie photos*

The selfie photo taking task allows participants to take a photo of themselves. It consists of picking up the phone from the table, moving it to the user to initiate, and then fine position the phone approximately arm's-length from the participant. A remote shutter button was used to take the picture once the phone was properly positioned. Fig. 6 shows a significant difference between the predefined and automatic positioning (*P = 0.035*) when using voice control interface. However, there was much less difference between positioning methods for gesture control.

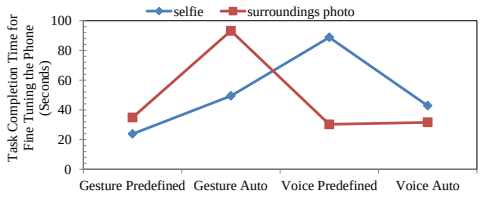


Figure 6. Average task completion time (fine positioning of the phone) for the selfie and picture taking task

IV. Discussion and Conclusion

In this paper, an integrated vision-based system is implemented to control a WMRM. The five modules of the system architecture worked in concert to enable persons with upper extremity mobility impairments to more accessibly and effectively perform a few common multistep tasks (i.e. drinking, making mobile phone calls and taking photographs). Two input modalities (gesture and voice recognition) were chosen for users to control the WMRM as they are accessible interfaces. A combination of HOG and SVM algorithms was applied to automatically detect and recognition the daily living objects. The participant's face and hands were recognized and tracked for automatic positioning of these objects for proper utilization. Feedback from a proximity sensor in the iPhone was used to facilitate the phone calling task. When the iPhone is placed near the ear, the WMRM automatically stops.

The pilot experimental results varied among tasks. For the straw drinking task verbal control using predefined positioning was the slowest but most accurate. There was not a strong correlation between slow completion time and less errors. There also were no correlations between these two factors when making a phone call. Likely, voice control tended to be slower than the other input modalities due to a delay in processing commands. For making phone calls and taking selfie photos, automatic positioning was significantly quicker and overall very accurate. We attribute the proximity sensor on the iPhone to assist in ideal placement to the user's ear. For selfie picture taking, the Kinect sensor was able to properly position the WMRM due to facial recognition programming. For taking photos of external objects at a distance, predefined positioning was quicker than automatic positioning. Voice control was also more efficient than gesturing for making minor adjustments to the camera screen.

Future work will consist of testing a combination of these input modalities. We will also recruit more participants to evaluate the usability of the presented system in real-world situations. In addition to performance, we will also evaluate participants' acceptance of these features.

References

[1] P. Schrock, F. Farelo, R. Alqasemi, and R. Dubey, "Design, simulation and testing of a new modular wheelchair mounted robotic arm to perform activities of daily living," in *IEEE International Conference on Rehabilitation Robotics, 2009. ICORR 2009*, 2009, pp. 518–523.

[2] S. L. Garber, A. L. Williams, K. F. Cook, and A. M. Koontz, "Article 14: Effect of a wheelchair-mounted robotic arm on functional outcomes in persons with spinal cord injury," *Arch. Phys. Med. Rehabil.*, vol. 84, no. 10, p. E3, Oct. 2003.

[3] H. Kwee, J. Quaedackers, E. van de Bool, L. Theeuwen, and L. Speth, "Adapting the control of the MANUS manipulator for persons with cerebral palsy: An exploratory study," *Technol. Disabil.*, vol. 14, no. 1, pp. 31–42, Jan. 2002.

[4] D.-J. Kim, Z. Wang, N. Paperno, and A. Behal, "System Design and Implementation of UCF-MANUS #x2014;An Intelligent Assistive Robotic Manipulator," *IEEEASME Trans. Mechatron.*, vol. 19, no. 1, pp. 225–237, Feb. 2014.

[5] W. G. Pence, F. Farelo, R. Alqasemi, Y. Sun, and R. Dubey, "Visual servoing control of a 9-DoF WMRA to perform ADL tasks," in *2012 IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 916–922.

[6] H. Tanaka, Y. Sumi, and Y. Matsumoto, "Assistive robotic arm autonomously bringing a cup to the mouth by face recognition," in *2010 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2010, pp. 34–39.

[7] R. Bloss, "Innovations like two arms, cheaper prices, easier programming, autonomous and collaborative operation are driving automation deployment in manufacturing and elsewhere," *Assem. Autom.*, vol. 33, no. 4, pp. 312–316, Sep. 2013.

[8] V. Maheu, J. Frappier, P. S. Archambault, and F. Routhier, "Evaluation of the JACO robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2011, pp. 1–5.

[9] I. Pathirage, K. Khokar, E. Klay, R. Alqasemi, and R. Dubey, "A vision based P300 Brain Computer Interface for grasping using a wheelchair-mounted robotic arm," in *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2013, pp. 188–193.

[10] A. Atrash, R. Kaplow, J. Villemure, R. West, H. Yamani, and J. Pineau, "Development and Validation of a Robust Speech Interface for Improved Human-Robot Interaction," *Int. J. Soc. Robot.*, vol. 1, no. 4, pp. 345–356, Nov. 2009.

[11] E. J. Rechy-Ramirez and H. Hu, "Flexible Bi-modal Control Modes for Hands-Free Operation of a Wheelchair by Head Movements and Facial Expressions," in *New Trends in Medical and Service Robots*, A. Rodić, D. Pisla, and H. Bleuler, Eds. Springer International Publishing, 2014, pp. 109–123.

[12] H. Jiang, J. P. Wachs, and B. S. Duerstock, "An optimized real-time hands gesture recognition based interface for individuals with upper-level spinal cord injuries," *J. Real-Time Image Process.*, pp. 1–14, 2013.

[13] K. M. Tsui, D. J. Feil-Seifer, M. J. Matarić, and H. A. Yanco, "Performance Evaluation Methods for Assistive Robotic Technology," in *Performance Evaluation and Benchmarking of Intelligent Systems*, R. Madhavan, E. Tunstel, and E. Messina, Eds. Springer US, 2009, pp. 41–66.

[14] H. Jiang, J. P. Wachs, M. Pendergast, and B. S. Duerstock, "3D joystick for robotic arm control by individuals with high level spinal cord injuries," in *2013 IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2013, pp. 1–6.

[15] H. Jiang, J. P. Wachs, and BS Duerstock, "Integrated vision-based system for efficient, semi-automated control of a robotic manipulator," *Intl. J. Intell. Comput. Cyber.*, pp. 253-266, 2014.

[16] P. Deléglise, Y. Esteve, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news," *Interspeech*, pp. 1653–1656, 2005.

[17] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working Set Selection Using Second Order Information for Training Support Vector Machines," *J Mach Learn Res*, vol. 6, pp. 1889–1918, Dec. 2005.

[18] https://itunes.apple.com/app/sensor-data-streamer/id608278214?mt=8