Simple Multimodal Object Detection and Classification for Brain Computer Interface via Kinect Sensor

Karthik Chellamuthu [#1], Guy Hotson [#2], Nathan E. Crone [*3], Nitish V. Thakor [#4]

[#] *Department of Biomedical Engineering. The Johns Hopkins University, Baltimore, Maryland.*
[*] *Department of Neurology. The Johns Hopkins University School of Medicine, Baltimore, Maryland.*
[1] kchella1@jhu.edu

***Abstract*—Control of dexterous robotic assistive devices necessitates advances in both neural decoding from a Brain Computer Interface (BCI) and in engineering paradigms that facilitate user interaction by exploring a variety of modalities. Among the latter, the Kinect-based computer vision system is vital towards monitoring a scene, tracking positions, and planning reaching and grasping tasks. In this work, we introduce the Kinect v2 as a novel, affordable sensor that is capable of improving the performance of the Hybrid Augmented Reality Multimodal Operation Neural Integration Environment (HARMONIE) in real-time by providing responsive yet accurate object detection. Image processing in conjunction with the use of linear binary classifiers proved to be unexpectedly effective and showcases the strength of the multimodal approach. The improved computer vision subsystem has the potential to offer more nuanced understanding of the environment to the modular prosthetic limb, (MPL) and ultimately to better serve persons afflicted by motor loss.**

I. Introduction

The Brain-Computer-Interface (BCI) is a hardware and software implementation that allows the user to interact with the surrounding environment in part or solely through cerebral activity. BCIs have come to the spotlight in recent years for being able to function as assistive technologies for patients experiencing motor loss or dysfunction. This includes patients paralyzed by neurological disorders such as amyotrophic lateral sclerosis, stroke, and spinal cord injury. Quadriplegics experiencing loss of motor function throughout their limbs cite upper-limb and hand functions as one of the most important contributors to their autonomy and quality of life [1]. It has been estimated that up to 230,000 people in the United States suffer from spinal cord injury which is only one subcategory of the affected demographic [2].

However, current BCIs are unable to give patients the level of neural control necessary to thoroughly restore upper-limb capabilities in a clinical setting [3], [4], [5], [6], [7]. It is therefore imperative to leverage advances in computer intelligence to improve system accuracy and offload some of the cognitive burden placed on the patient. To begin addressing these challenges, we have worked towards building a hybrid interface model that is capable of operating semi-autonomously using multiple inputs including manual control, eye tracking, EEG, and sEMG to control the JHU/APL Modular Prosthetic Limb (MPL) [8].

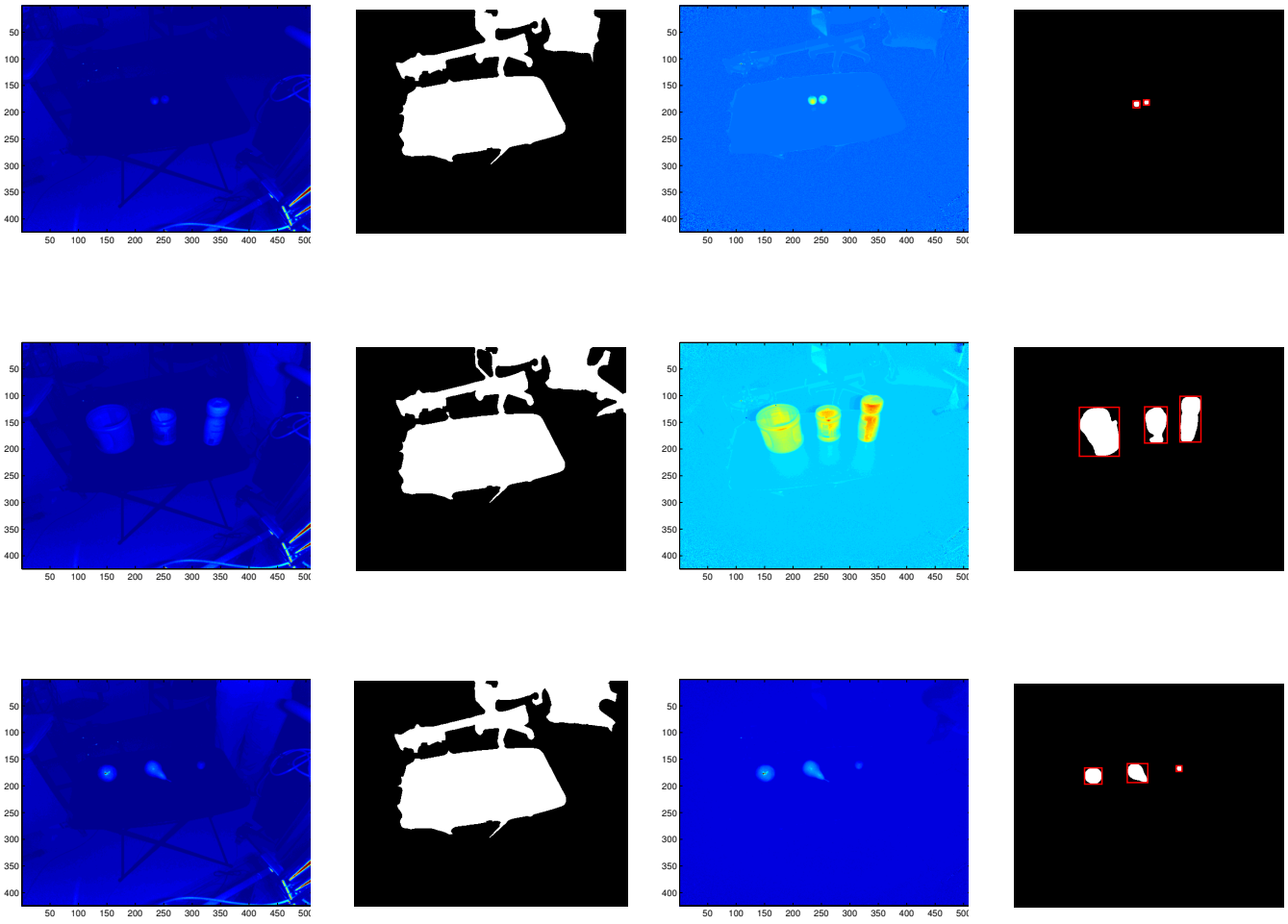
Furthermore, due to advances in sensing, hardware, and algorithm design, computer vision has found applications in real-time rehabilitative robotics systems [9], [10], [11]. The implementation of such a subsystem, which will be the focus of this work, enables scene reconstruction, object recognition, and model generation. In the aforementioned Hybrid Augmented Reality Multimodal Operation Neural Integration Environment (HARMONIE), the original Microsoft Kinect sensor was used to perform scene reconstruction using point-cloud conversion. However, due to the noisy nature of the structured lighting method used to reconstruct the image, the original sensor could not reliably classify and track objects [12].

In contrast, the Kinect v2 not only has a larger field of view, but also a fundamentally different mechanism for depth estimation. Unlike the previous model which attempts to structure a projected speckle pattern of infrared laser, time of flight (ToF) sensing depends much less on lighting based calculations, behaving analogously to sonar which calculates distance traveled from the phase shift for the signal to return to the sensor [13]. This approach is also better suited for real-time applications due to direct computation of all points in the scene instead of interpolation of non-lighted regions. In conjunction with the high contrast infrared (IR) feed, the Kinect v2 computer vision system allows for accurate object tracking and is shown here to be viable for incorporation within the hybrid BCI framework.

II. Methods

A. Overview

In the original HARMONIE system, objects are placed on a table measuring 66 by 45 centimeters, approximately half a meter away from the Kinect. The modular prosthetic limb was then placed underneath the Kinect at a resting position in order to later perform grasp tasks. For this study, an identical setup was used but with a newer model of the Kinect. The latest model comes equipped with Active IR video feed for



(a) *Original IR Image.* (b) *Region Growing Mask.* (c) *Background Subtracted Image.* (d) *Final Segmented Image.*

Fig. 1: Schematic overview of object segmentation process. The raw infrared data from the Kinect (a) is used to approximate the workspace using pixel intensity based region growing (b). The resulting mask is then subtracted from the original to yield (c) after rescaling. Filtering, preprocessing, and edge detection generate the final image (d).

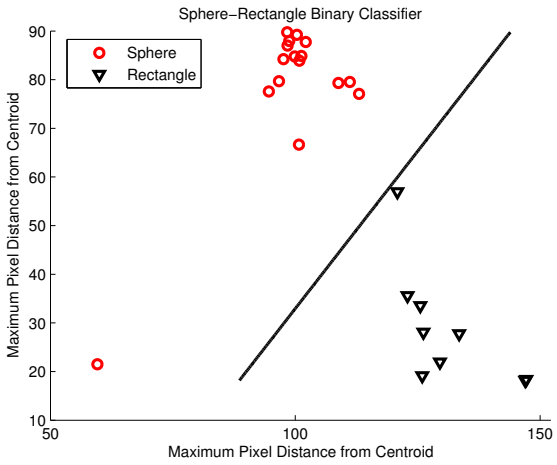
dark/low lighting conditions, 1920x1080 resolution, and a Time-of-Flight depth sensor [13]. At the current time, the Kinect provides no direct interface to MATLAB in which the majority of the processing occurred. Thus, the raw data from the Kinect sensor was outputted in binary format and read in real-time into MATLAB with negligible latency. Data could be read from disk and streamed at an average rate of 0.02 sec/frame. Once both the IR and Depth images have been processed, the extracted information is streamed to the Robot Operating System (ROS) using a User Datagram Protocol (UDP) interface.

B. Workspace Segmentation

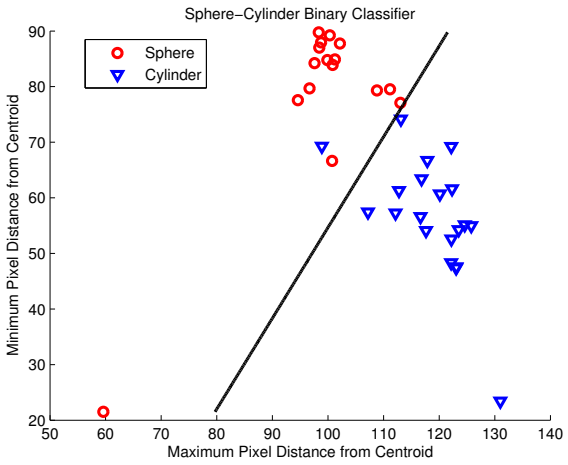
At startup, the Kinect first attempts segment the table workspace by creating a pixel intensity mask which is later used to eliminate unwanted background. Since the Active IR stream is heavily dependent on material type and less so on lighting, the workspace to be identified will be displayed as a contiguous area which is suited to region-growing schemes [14]. The particular algorithm implemented begins at a seed point which by default is the center of the image but can be adjusted if needed. From here, the neighboring eight pixels are assessed to find the one which minimizes the difference between the average intensity of the region and that of the pixel in question. If the final difference is less than a predetermined threshold, then the region will grow to encompass that pixel. This process is then called recursively on the newest pixel of the region and terminates when no other qualified pixels can be found. Once the region has been identified, small gaps created by the objects on the workspace will be filled to construct a final mask. During testing, the region growing approach was able to correctly segment the table in 46/50 trials.

C. Object Segmentation

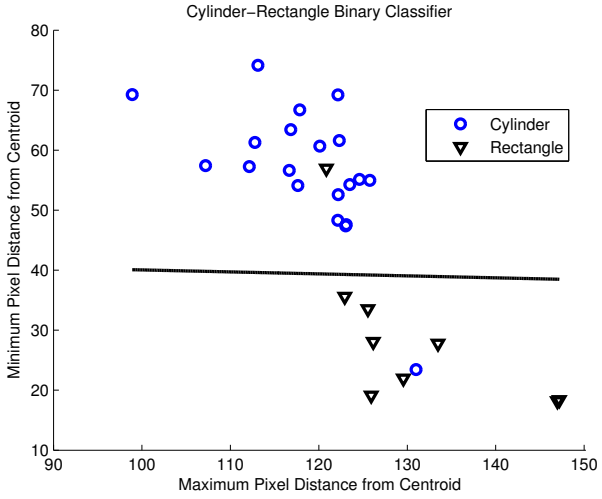
Once a reliable segmented background can be created, the resulting mask is subtracted from the incoming IR stream



(a) Classifier 1



(b) Classifier 2



(c) Classifier 3

Fig. 2: Plots of training data for three binary classifiers suggest categories are linearly separable. LDA function is plotted along with data for each scenario ((a) Sphere-Rectangle, (b) Sphere-Cylinder, (c) Cylinder-Rectangle).

in real-time. This subtracted image undergoes thresholding and further preprocessing to remove subtraction artifacts and remaining noise. To properly discriminate objects and eliminate misleading high frequency spatial information, a Gaussian low-pass filter is applied. From here, Canny Edge Detection [15] is utilized to mark bounded object regions, completing the segmentation phase. Parameters, including threshold percentile, kernel values, and step differences, were tuned offline using histogram analysis of the IR intensity values and by visual inspection. Figure 1 summarizes this process for closely neighboring, cylindrical, and spherical objects.

D. Object Classification and Tracking

1) Shape Classification: Once an object in a scene is segmented, the rough shape and relative distance from the Kinect must be determined and relayed to the prosthetic limb to plan grasp type and approach trajectory. One prominent feature used for classification is the relationship between the minimum and maximum distances from the boundary to the centroid of the object. That is, for an object with centroid $[C_x\, C_y]^\mathsf{T}$ and boundary $\mathbf{B}$, we have a feature vector $\mathbf{v} = [v_1\, v_2]^\mathsf{T}$ where

$$v_1 = \min_{\forall P_i \in \mathbf{B}} \sqrt{(C_x - x_i)^2 + (C_y - y_i)^2}$$

$$v_2 = \max_{\forall P_i \in \mathbf{B}} \sqrt{(C_x - x_i)^2 + (C_y - y_i)^2}$$

and P_i is an arbitrary point along $\mathbf{B}$ with coordinates (x_i, y_i). To ensure equivalent relative distances each object was normalized to a size of 100x100 pixels prior to calculating the feature vector. Plotting the labeled data reveals that the shapes (e.g. cylinder, sphere, and rectangle) can be linearly separated (see Figure 2). Thus, Linear Discriminant Analysis (LDA) was chosen as a viable classifier.

LDA is a supervised learning method that seeks to find the linear combination of features that best separates two data sets. Assuming a normally distributed spread in both classes, this means the selected line will maximize the distance between the two means while simultaneously minimizing the individual variances [16]. In this study, 3 LDAs were trained on a training set of 45 unique instances of objects with varying shape, position, and orientation. The resulting classifiers were then implemented in real-time.

2) Range Estimation: In order to track the position of a given object within the scene, the Time-of-Flight (ToF) depth sensor is used in conjunction with the objects pixel position in the IR feed to calculate a 3D Euclidean distance. The point cloud distance is computed by first converting each dimension in the image and depth to scaled values and finding the magnitude of this vector.

III. Results

A. Segmentation Results

During online testing, the computer vision system was able to segment and identify two spheres of radius 0.4 inches placed as close as 0.75 inches apart. Cylindrical objects with a radius ranging from 2.125-2.6 inches and height from 4-7.25 inches were also successfully segmented and identified. The segmentation algorithm was tested in 30 trials each with objects of varying shape, orientation, and material composition. All objects were properly segmented in 26 of these trials and up to eight objects could be simultaneously segmented and tracked.

B. LDA Classifiers

1) Classifier 1: Average classification accuracy for the sphere-rectangle binary classifier after 10-fold cross validation was $93.5 \pm 1.69\%$.

2) Classifier 2: Average classification accuracy for the sphere-cylinder binary classifier after 10-fold cross validation was $94.0 \pm 1.37\%$.

3) Classifier 3: Average classification accuracy for the cylinder-rectangle binary classifier after 10-fold cross validation was $96.5 \pm 1.28\%$

C. Overall System Performance

The computer vision system was benchmarked both online and offline to assess performance. To gauge expected runtime and latency values, the entire routine has been broken down into stages as shown in Table I. In addition, the distance measurements taken from point cloud conversion of the depth images were analyzed. These values were found to range from 0.5105 to 9.5438 point cloud units over 100 frames with a variance of 5.468×10^{-6}.

Latency Table	Average Time (sec/frame)	Standard Error (sec/frame)
Region Growing	0.8012	0.0136
Segmentation	0.1302	1.7989×10^{-6}
Classification	0.0103	3.8754×10^{-4}

TABLE I: System Performance Measurements

IV. Discussion

In this paper, we detailed the implementation of an improved computer automated object recognition and tracking system which is compatible with the existing HARMONIE architecture. The new Active IR feed and depth sensor in the updated Kinect provide better image contrast and stability, enabling natural application in human-computer interaction systems. The results obtained from utilizing this technology show improved classification accuracy and less image noise compared to the previous Kinect system. Moreover, the performance and responsiveness of this system make it feasible to implement in a real-time setting. While the current system is tailored for classification through simple shape descriptors, this approach can be expanded to include more sophisticated 3D reconstruction algorithms that can precisely account for occlusion, texture, and volumetric parameters for the purpose of finding optimal strategies for object manipulation by the MPL hand.

Acknowledgment

The authors would like to thank the Microsoft Kinect Developer Program for early access to the Kinect for Windows v2. We also acknowledge Brock A. Wester for helping to fabricate a new stand for the Kinect.

References

[1] C. Zickler, V. Donna, V. Kaiser, Al-Khodairy, S. Kleih, A. Kubler, and M. Malavasi, "Bci applications for people with disabilities: Defining user needs and user requirements." *AAATE 25th Conference, Florence, Italy August-Sept*, pp. 185–189, Sept 2009.

[2] A. I. Nobunaga, B. K. Go, and R. B. Karunas, "Recent demographic and injury trends in people served by the model spinal cord injury care systems," *Archives of Physical Medicine and Rehabilitation*, vol. 80, no. 11, pp. 1372–1382, Nov 1999, lR: 20061115; JID: 2985158R; ppublish.

[3] V. Gilja, P. Nuyujukian, C. A. Chestek, J. P. Cunningham, B. M. Yu, J. M. Fan, M. M. Churchland, M. T. Kaufman, J. C. Kao, S. I. Ryu, and K. V. Shenoy, "A high-performance neural prosthesis enabled by control algorithm design," *Nature neuroscience*, vol. 15, no. 12, pp. 1752–1757, Dec 2012, lR: 20141104;.

[4] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia," *Lancet*, vol. 381, no. 9866, pp. 557–564, Feb 16 2013, lR: 20141104; CI: Copyright (c) 2013;.

[5] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, May 16 2012, lR: 20141016;.

[6] T. Yanagisawa, M. Hirata, Y. Saitoh, H. Kishima, K. Matsushita, T. Goto, R. Fukuma, H. Yokoi, Y. Kamitani, and T. Yoshimine, "Electrocorticographic control of a prosthetic arm in paralyzed patients," *Annals of Neurology*, vol. 71, no. 3, pp. 353–361, Mar 2012, cI: Copyright (c) 2011;.

[7] M. S. Fifer, G. Hotson, B. A. Wester, D. P. McMullen, Y. Wang, M. S. Johannes, K. D. Katyal, J. B. Helder, M. P. Para, R. J. Vogelstein, W. S. Anderson, N. V. Thakor, and N. E. Crone, "Simultaneous neural control of simple reaching and grasping with the modular prosthetic limb using intracranial eeg," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 22, no. 3, pp. 695–705, May 2014, gR: 3R01NS0405956-09S1/NS/NINDS NIH HHS/United States;.

[8] K. Katyal, M. Johannes, T. McGee, A. Harris, R. Armiger, A. Firpi, D. McMullen, G. Hotson, M. Fifer, N. Crone, R. Vogelstein, and B. Wester, "Harmonie: A multimodal control framework for human assistive robotics," *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pp. 1274–1278, Nov 2013, cI: Copyright (c) 2011; JID: 7707449; 2011/02/04 [received]; 2011/08/04 [revised]; 2011/08/12 [accepted]; 2011/11/02 [aheadofprint]; ppublish.

[9] H. Jiang, J. P. Wachs, and B. S. Duerstock, "Integrated vision-based robotic arm interface for operators with upper limb mobility impairments," *IEEE ...International Conference on Rehabilitation Robotics : [proceedings]*, vol. 2013, p. 6650447, Jun 2013.

[10] A. Frisoli, C. Loconsole, D. Leonardis, F. Banno, M. Barsotti, C. Chisari, and M. Bergamasco, "A new gaze-bci-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 1169–1179, Nov 2012.

[11] B. Choi and S. Jo, "A low-cost eeg system-based hybrid brain-computer interface for humanoid robot navigation and recognition," *PloS one*, vol. 8, no. 9, p. e74583, Sep 4 2013, lR: 20141112; JID: 101285081; OID: NLM: PMC3762758; 2013 [ecollection]; 2013/06/13 [received]; 2013/08/06 [accepted]; 2013/09/04 [epublish]; epublish.

[12] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1318–1334, Oct 2013.

[13] M. Corporation, "Kinect for windows v2 features," 2014, `http://microsoft.com/en-us/library/dn782025.aspx`.

[14] J. Fan, D. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 10, no. 10, pp. 1454–1466, 2001, jID: 9886191; ppublish.

[15] J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, no. 6, pp. 679–698, Nov 1986.

[16] R. A. FISHER, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, Sept 1936.