Hover: A Wearable Object Identification System for Audio Augmented Reality Interactions

Augmented reality (AR) has recently begun to gain popularity and become accessible to mainstream audiences. Companies such as Microsoft, Google, and Magic Leap are developing platforms that bring augmented reality to the masses through smartphones and head-mounted devices. This technology has many potential uses, some promising ones being guiding users through tasks that they may not be experts in [3], or overlaying helpful information on objects[4]. In these cases, AR systems provide access to an object's contextual information- or additional information linked to an object in the physical world. However, even with all the advances of today's AR systems, users generally report that these are too intrusive, distracting, and cumbersome for daily use [1,2]. As an example, a phone-based AR application that uses the display to overlay information about an object, needs the phone to be between the user and an object of interest. This can distract a user from his or her surroundings, obstruct a user's field of view, and potentially interrupt them on whatever activity they were doing beforehand (if the phone is not in an easily accessible place). On the other hand, the glasses form factor, such as the one that the Hololens has, alleviates some of these problems by putting the contextual information directly in a person's field of view at any time. However, the person always needs to carry the typically bulky headset with them to have access to the information. Adding to this, the field of view of current devices is small. Lastly, apart from focusing on these form factors, most of the current systems do not take advantage of sensors other than cameras, which can adversely affect their object detection performance in varied lighting and with visually similar objects.

Based on the existing ecosystem, we frame the following question: *Can we find a multi-modal and seamless way to access contextual information tied to objects in our environment?* To explore this question, we design, implement, and test a sensor-fusion powered wearable object identification system which allows users to "hover" their hands over objects of interest and access contextual information that is tied to the object through voice interactions with an intelligent assistant (we use Mycroft[9] and custom developed skills, with a bone conduction headset for discreet communication). We wanted the system to give discreet access to AR contextual information without the need of phones or headsets. We called this system Hover.

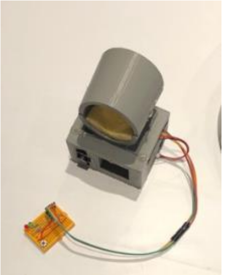

Figure 1 -
Prototype for the Hover system.

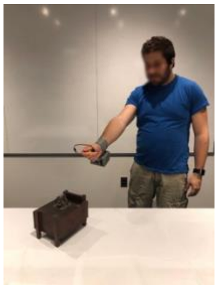

Figure 2
Use case: user is scanning an item with the Hover system on his wrist and a bone conduction headset.

Hover is a unique system that uses a fusion of sensors to perform the identification of an object under a variety of conditions and is capable of discerning between visually similar objects. Among these sensors there is a camera (operating in the visible and infrared spectrum), a small solid-state RADAR, and multi-spectral (visible and near infra-red spectrum) light spectroscopy sensors. In our system we extract features from each sensing modality, normalize them according to their sensor of origin and combine them into one feature vector. We then pass this feature vector into a classifier. Some of the features we extracted were Bag of Visual Words histograms[10] for images, High Resolution Range Profiles[12] and Range-Doppler maps[11] for the radar, and the root mean square of the readings for the spectrometers. For the classifier we evaluated a Support Vector Machine[5], a Random Forest[6,7], and a K-Nearest Neighbors[8] to see which could deal more easily with classifying the multi-modal data. We found that without any optimization the Random Forest classifier (RFC) gives us the best performance of the lot. Utilizing the RFC, we found that the Hover's sensor fusion system is slightly better (78% vs. 75% accuracy) than a plain camera without any sort of optimization in a classification task of 9 objects.

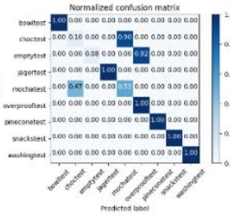

Figure 3
Base Image System
Confusion Matrix

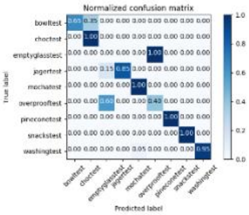

Figure 4
Fusion System
Confusion Matrix

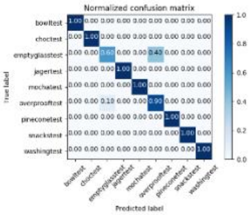

Figure 5
Fusion System
without NIR
Spectrometer

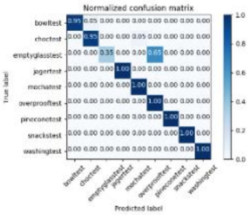

Figure 6
Fusion System
without RADAR

We note that the performance of the system could possibly be improved if we replaced the RADAR module with one that has more bandwidth and range resolution, or we evaluated different RADAR features. The features we selected, in combination with the RADAR hardware we utilized, introduced confusion into the classification as can be seen in figure 6. Additionally, we note that the performance could be improved further if we install a broadband Infra-Red (IR) LED for our Near IR Spectrometer, rather than the one used mistakenly in testing, which operates in one specific frequency rather than throughout the range the sensor reads, and therefore also introduces confusion into the system. Additionally, we are currently exploring the possibility of inputting the raw sensor fusion signals into a neural network architecture to see if learned features, rather than hand crafted ones, could improve the overall performance of the system.

With regards to the user interaction aspect, we developed an identification task in which we placed participants in a predetermined environment to identify a set of objects with a certain device. Once the device identified the object it proceeded to give information about it through an assistant in the device. Participants were asked to repeat the identification task with Hover, Hololens, and an Android phone. The task simulated a situation where a user needs to gather information, like in a museum or while shopping. Participants were asked a set of questions and we collected answers in a 5-point Likert scale, to gage their individual and overall experience with each device. After analyzing the results, we found that Hover is comparably cumbersome with a mobile phone, and less cumbersome than the head-worn Hololens. Also, Hover is comparably immersive to the Hololens. Hover is also individually less interruptive than a phone but was perceived as just as interruptive as all the other devices in the study. It is comparable in intrusiveness with a mobile phone, and that it is less intrusive than the head-worn Hololens.

We have identified key areas of improvement for future work. The most prevalent is the form factor, which could be improved by streamlining the electronics into a smaller integrated board, shrinking the overall footprint. Shrinking the size of the device would improve the overall user experience as the device would conform to a wrist and feel more natural. To make it easier for new users, we also will attach a laser pointer so that users can see which objects are in Hover's frame of view. Another component that could be improved in future iterations is the intelligent assistant. Currently, the Intelligent Assistant only gives the contextual information that was given when training an object unto the system. We are looking into how to make dynamic information retrieval about a topic/object and how to process it and make inferences from it to answer pertinent questions or doubts that users may have. We intend to make this component the focus of our future research, as an assistant with dynamic knowledge acquisition, could be utilized in a variety of activities such as interactive grocery shopping, interactive dieting systems, and interactive usage tutorials based on a scanned device of interest.

References:

[1] Velamkayala, Eswara Rao, Manuel V. Zambrano, and Huiyang Li. "Effects of HoloLens in Collaboration: A Case in Navigation Tasks." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 61. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2017.

[2] Van Krevelen, D. W. F., and Ronald Poelman. "A survey of augmented reality technologies, applications and limitations." *International journal of virtual reality* 9.2 (2010): 1.

[3] Westerfield, Giles, Antonija Mitrovic, and Mark Billinghurst. "Intelligent augmented reality training for motherboard assembly." *International Journal of Artificial Intelligence in Education* 25.1 (2015): 157-172.

[4] Tang, Arthur, et al. "Comparative effectiveness of augmented reality in object assembly." Proceedings of *the SIGCHI conference on Human factors in computing systems*. ACM, 2003.

[5] Hearst, Marti A., et al. "Support vector machines." *IEEE Intelligent Systems and their applications* 13.4 (1998): 18-28.

[6] Gilles Louppe. 2014. Understanding random forests: From theory to practice. *arXiv preprint* arXiv:1407.7502(2014)

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008.*The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

[8] Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia* 4, 2 (2009),1883

[9] Mycroft - open source artificial intelligence for everyone. Available at
https://mycroft.ai/.

[10] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski,and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision*, ECCV, Vol. 1.Prague, 1–2

[11] Xuejun Liao, Paul Runkle, and Lawrence Carin. 2002. Identification of ground targets from sequential high-range-resolution radar signatures. *IEEE Transactions on Aerospace and Electronic systems* 38, 4 (2002),1230–1242

[12] Andrzej Wojtkiewicz, Jacek Misiurewicz, Konrad Jedrzejewski, and Krzysztof Kulpa. [n. d.]. Two-dimensional signal processing in FMCW radars. ([n. d.]).