

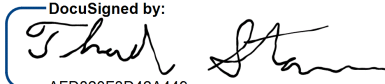
Artificial Intelligence (AI) through Symbiosis

Devansh Jatin Ponda

Georgia Institute of Technology

Faculty Member #1:

Dr. Thad Starner

Signature: _____

AFD820E3D42A449

Faculty Member #2:

Dr. Gregory Abowd

Signature: _____

801392E17071496...

Table of Contents

1. ABSTRACT	3
2. INTRODUCTION	4
3. LITERATURE REVIEW	7
4. METHODOLOGY	11
4.1. OVERVIEW	11
4.2. SETUP AND DATA COLLECTION	12
4.3. EXTRACTING IMAGE OF HAND/PALM FROM COLLECTED VIDEO	13
4.4. BUILDING THE MODEL	14
5. RESULTS	16
6. DISCUSSION AND FUTURE WORK	18
7. CONCLUSION	20
8. CITATIONS	21

1. Abstract

While most traditional machine learning approaches tend to focus more on data – cleaning, filtering, and formatting data – I ask if we could instead focus more of our attention on trying to make our computers understand the nuances of the real-world instead. What if instead of humans feeding cleaned data to computers, there could be a ‘symbiosis’ in which the computer actively learns the habits of humans through their everyday interactions with the world, and assist them in these interactions?

In my research, I have tried to build one of the earliest systems that follow the AI-through-Symbiosis framework in the context of order picking. By collecting video data from a head-mounted GoPro camera that a user wears while picking items, I try to recognize – without any data labelling – which item the user has picked. The idea is that if this information is known, we can then use it to determine if the user has picked an incorrect item and alert them, which would in turn help reduce the error rates in the process.

In this thesis, we explore the idea of AI through Symbiosis, the motivation to study it, the methodology of collecting data in the context of order picking, preliminary results, and possible future work.

2. Introduction

Humans and computers – the interaction between the two has been a driving force in everything ranging from landing the first man on the moon about 50 years ago [1], to communicating across the world in real-time today.

While for most of history the relationship between the two has been one where humans are actively involved in teaching the computer exactly what to do through some form of computer code and the computer passively executes these instructions, the past decade has seen a seismic shift in this approach. In recent years, artificial intelligence has led to a novel way to teach a computer where it learns to draw patterns from data and assists decision making after completion of initial training. While the computer's role has shifted from being a passive 'executor' to an active learner, the role of the human is still that of an actively involved teacher, because it is the humans that create clean, consumable data for the computer to analyze and hence train itself.

However, what if there was a way in which the computer's role remains that of an active learner, but we humans can passively sit back and go on with our lives, while the computer learns from our actions? What if the relationship between humans and computers can be deepened to an extent where the fine line separating

the two gets blurred? This is the Artificial Intelligence through Symbiosis (or AI through Symbiosis, or Symbiotic AI) approach that this study examines.

Artificial Intelligence (AI) is the branch of computer and information sciences that focuses on developing hardware and software systems that solve problems and tasks that if accomplished by humans would be considered a display of intelligence [2]. Machine Learning (ML) is a sub-field of AI that enables systems to learn automatically from big data without being explicitly programmed to accomplish a task a certain way.

Artificial Intelligence through Symbiosis is the idea of having a computer learn a user's behaviors from their everyday experiences and eventually assist the user in carrying out these tasks. My research will focus on building a Symbiotic AI system in the context of order picking.

Order picking is the process of retrieving products from storage (for example, a warehouse) in response to a specific customer's request [17]. From the 750,000+ warehouses around the world, over a trillion dollars' worth of goods are annually picked [3]. With the large-scale operations in this field, it is clear that any optimization that can improve pick speed or accuracy can potentially save businesses millions, if not billions of dollars each year.

The status quo for order picking operations is paper-based pick lists that have information about the item, such as its name, description, and location in the warehouse. Workers pick out items order by order and drop them off at specified locations. It has previously been determined that head-worn displays, such as the Google Glass or the Magic Leap One, generally perform better than paper for order picking because they not only enable the worker to pick with both hands, but they also make it easier to determine the closest next item to be picked [5]. However, little work has been done in using the visual data that can potentially be generated from these devices to optimize the process.

To further study how the video data generated from head-worn displays can be used, this study aims to train ML-based algorithms to determine whether or not the user has picked the correct item and alert them immediately to swap the wrong item for the correct one. It is predicted that the analysis will show a significant drop in the error rates per order picked when compared to the industry standard: paper.

3. Literature Review

Artificial Intelligence through Symbiosis is a very new, promising field in computer science. While there has been speculation as to what the motivations are, what an end-goal system could look like, and how one can bring artificial intelligence and humans together (as explained further), almost no work has been done that empirically examines an AI through Symbiosis approach in a real-world setting.

In his talk titled ‘Symbiotic AI: Creating an AI with Human Level Experience’ [6], Dr. Thad Starner talks about the problems with building a system, like a robot, embedded in the stochastic, real world, and teach it to navigate a room using traditional AI approaches of using clean and ‘curated’ data. He argues that because our everyday experiences are not filled with such clean data made specifically for machine learning jobs, we should not use this data to have a computer ‘learn about the human world.’ What can be derived from this talk is the need for better approaches where our systems are able to understand what we humans are specifically focusing on in the real-world and learn to differentiate those entities from all the noise in the environment. We need to teach them to not only understand our behaviors with specific items in our environment, but also find a way for them to learn what those items are and separate them from the rest of the

environment. This need to help computers understand the nuances of the real-world is one of the main reasons why the AI through Symbiosis approach must be developed.

An unrelated but interesting outlook towards bridging the gap between humans and AI can be found in ‘Symbiotic Artificial Intelligence’ [7] by Pramod Kotipalli. He states: ‘The first, and perhaps most important, principle of Symbiotic AI is of reducing the time (and effort) between intention and action. Once the time between intent and action reduces below two seconds, we are able to start forming habits.’ This is important because any eventual system that truly embodies the principle of AI through Symbiosis should be able to learn patterns in the user’s behavior and predict, based on past experiences, what the user would like to do next and assist them in doing so. Building this ‘real-time’ system will require heavy computational capabilities as close as possible to the device that the user is wearing. While this study will focus on building an offline prototype of a system that obeys AI through Symbiosis, the idea that AI through Symbiosis can be best achieved in a real-time system is an important one to note.

This study explores using Artificial Intelligence through Symbiosis in the context of order picking. The tools and techniques used to build the software system are those that build off of previous work in computer vision. Thus, while much of the

work in AI through Symbiosis has been speculative, the variety of different work done in both order picking and computer vision that this study uses, is important to discuss.

In the last decade alone, there have been several different works that have explored possible ways to optimize the process of order picking. It has been previously shown that using head-worn displays, instead of paper-based pick lists, improves the efficiency of the order picking process [5]. An important consideration is how comfortable workers are while wearing these devices over extended periods of time and how this comfort can be improved. It has also been previously shown that visual information displayed between 0 and 20 degrees from the center of the field of view is much easier for users to quickly read and understand when compared to information displayed beyond 20 degrees from the center of the field of view [8].

Even in the computer vision space, a lot of work has been done in segmenting human features like faces, hands, etc. from video/image data [9, 10, 11, 12]. Work has also been done in edge detection techniques that can potentially be used to differentiate between a foreground and a background in an image. Two famous techniques for achieving this are Sobel filters and Canny filters [13, 14]. These filters essentially use the gradient of pixel values in the image to determine where the edge of an object could possibly lie – where the pixel values change abruptly

there may be a transition from the foreground to the background. All of these works are building blocks that can together be used in the system powered by AI through Symbiosis that this study builds.

4. Methodology

4.1. Overview

There were three parts to how this project was conducted and how it can be replicated. The first being setting up the lab to collect data; the second was designing and building the software systems and processing the collected data through a part of this system; finally, building the machine learning model and testing the accuracy of this classifier.

At a high-level, we have built an offline (not real-time) software pipeline that can automatically (without any manual data labelling) identify what item the user has picked.

To do this, we process data coming in from a head-worn camera, the ‘GoPro Hero 7 4K’, that a user wears while carrying out the order picking operation. In an industrial setting, this video feed would come from the user wearing a head-worn display, like the Google Glass, where they can also see visual information about which item has to be picked. We used the GoPro instead of the Google Glass in this study because of how much more flexibility the GoPro gives us in terms of configuring the field of view. This is useful for our study because it allows us to direct the line of sight of the GoPro diagonally downwards where the user’s hand would be, thus allowing us to capture video data.

4.2. Setup and data collection

Because the idea was to build a proof-of-concept, we established a very simple lab design. The setup included source bins, target bins, and objects. We used two different classes of objects with different colors: red and green. Within each class, all objects used were identical. We used items with different colors so that there is a sharp contrast between the colors which were hypothesized to make it easier for future stages of the pipeline to build the classifier that can identify if the user has picked the correct item.

The source bins and target bins were kept at opposite ends of the lab. The source bins consisted of several of the items from these different classes. The user, after head-mounting the GoPro, conducted several picks for data collection. Here, a pick is defined as the process of picking one item using one hand from one of the source bins, walking to the target bins with the item still held in the palm of the open hand, dropping the item into one of the target bins, and returning with an open, empty hand back to the source bins. For each of the different classes of objects, there were at least 10 picks. This minimum number of picks was conducted to ensure that we are able to collect enough data for subsequent steps in the system.

4.3. Extracting image of hand/palm from collected video

After collecting the data, we input it into the first step of the software system. We built this system using Google MediaPipe [12]. Mediapipe is an open source library in C++ built by Google to help support computer vision tasks like hand recognition or face tracking in image/video input files. We chose MediaPipe because it had high accuracy in hand detection, and that it was completely open-source, which allowed us to modify the source code to match our own project needs. We used MediaPipe to track the motion of the hand in the field of view and modified the code to extract the image within the bounding box of the hand. A bounding box is essentially the ‘region of interest’ we are trying to extract. The image within this bounding box is then used, through MediaPipe, to extract the image of the palm. Just like with the image of the hand, the image of the palm is also presented by MediaPipe in the form of a bounding box, whose coordinates we can then use to crop it out programmatically.

Because MediaPipe is not perfect, and sometimes has false positives (regions from the frame that do not contain a hand, and yet MediaPipe identifies it to be a hand), we used the confidence score presented by MediaPipe to filter out all the frames in which MediaPipe was less than 95% confident that the frame contained an image of a hand.

4.4. Building the Model

With the data collected and processed from section 4.3, we built a machine learning model that could classify the images. However, it was observed that not all images processed were of the same size – there are significant differences based on how far the hand was from the camera. The smallest image observed in our dataset was found to 40x40 pixels. So, we downsized all other images to this size so that we can use it for machine learning.

The machine learning approach we use is called K-Means clustering [15]. This technique groups together data items that are similar to each other in K number of different groups, where we specify the value of K. The value of K is 1 more than the number of different objects in the dataset. For example, if we had 3 different types of items being picked, then $K = 4$ because the first three groups would be for the images of items themselves, and the last group would be for images of an empty hand. The key reason we used K-Means clustering is because we want to use an ‘unsupervised’ approach to machine learning which doesn’t require any data labelling. Data labelling is the ‘process of detecting and tagging data samples’ [16]. Data labelling is an important requirement if a ‘supervised’ machine learning approach has to be used, and would serve as a significant utilizer of both the

resources of time and money for businesses. An advantage to using unsupervised learning like we did, is that an industry-scale implementation is easier.

5. Results

Experiments, following the guidelines stated in the Methodology section, were conducted to collect data. The data was collected in the form of videos of the order picking operation.

Each video consisted of at least 10 picks. Upon completion of 10 picks for each class of objects, 2 videos – each about 3 minutes in length – were generated. This data was then fed into our system.

The first step in the system, as described earlier, is running the videos through Google MediaPipe [12], which would extract the image of the hand from each frame, and then extract the image of the palm (which may or may not contain an object). Processed image results extracted from this stage of the system are shown below. Overall, upon completion of this first step in the system, we extracted about 4500 images for each 3-minute video, leading to a cumulative dataset of $4500 \times 2 = 9000$ images.



Processing the raw video produces an Image of the Hand



Processing the Image of the Hand produces an Image of the Palm

Running the next stage of the system converts the image into raw pixel values, downsizes it to be the size of the smallest image, and then gets pushed into a K-Means cluster for classification. Upon manual calculation, it was discovered that for each of the categories – red object in hand, green object in hand, and no object in hand – we had accuracies of about 54%, 71%, and 91% respectively for identifying which object is in the hand.

6. Discussion and Future Work

The system has yielded varying levels of results at different stages. For the first stage – that of extracting the image of the hand and the image of the palm – the results are very good – possibly as good as they can get using an approach based on visible light. The only issue was that roughly 9% of the images were ‘bad’ data, because they neither showed an empty palm, nor did they show a palm with an object in it. Instead, these images were extracted out of the data captured while an item was being picked from the source bin or being dropped into the target bin. Thus, the contents are often the side view of the hand or that of the back of the hand, neither of which helps the machine learning classifier. Since we could not figure out any other way to filter these images out except by doing it manually, we decided to keep them in the dataset given it only affected 9% of the images. Future approaches to this stage of the pipeline could explore using infrared light to segment the image of the object within the hand. This technique would rely on the assumption that the object being picked has a significantly different temperature than that of normal human body temperature of about 37 C or 98.6 F. Because of this temperature gradient and because of how useful infrared cameras are at detecting this temperature gradient, filters (like Canny filters or Sobel filters [13, 14]) could be used to segment the image of the object/palm from that of the hand. These filters work by segmenting an image near regions of high gradient or

change, and since our key assumption is that the temperature difference between the hand and the object is sharp (or at least sharper than gradients within the hand or within the object), filtering would possibly be quite accurate.

For the second stage of the system, the results were great for a first attempt in this promising area in computer science and AI. One of the reasons why the accuracy of the classifier was not higher is possibly because downsizing images led to very few differences between distinct images at the pixel scale. When the first stage is run, depending on how far the hand is from the camera, the size of the extracted image of the palm can be as little as 40x40. Other images, which are significantly larger in size, thus have to be downsized by averaging the neighboring pixel values around a pixel, thus losing information. An approach that can be explored in the future would be to set a certain minimum threshold for image size, such that all images sized below this threshold would be ignored. This would ensure that all images are not completely downsized because of a few outliers and could potentially lead to better results because information is not lost. However, a tradeoff in this approach would be to risk not being able to use all the collected data and could lead to more data needing to be collected which would increase the time needed to implement the system at scale.

7. Conclusion

The two main goals of this thesis were to build a technological proof-of-concept of the AI through Symbiosis approach and to establish the potentially large scale, high monetary impact of using this revolutionary new approach to building AI-at-scale, both of which it has done.

Technologically, we have built an offline (i.e. not real-time) system in the specific context of order picking, but the results we derive from this are generalizable to other industrial and non-industrial applications as well. There are several avenues for improvement of the current solution using new and innovative techniques to all stages of the system.

Additionally, we have produced a high-quality dataset open for others to use.

Finally, we have shown a new, non-traditional approach to building AI and we hope this project inspires the wider scientific community to use our dataset and build on the existing framework. With efforts from researchers and scientists from all across the world, we hope that we have triggered the process of building the smartest, most personalized, and most convenient artificially intelligent assistant that can improve the human condition.

8. Citations

- [1] "Apollo Guidance Computer." *Wikipedia*, Wikimedia Foundation, 4 Nov. 2020, en.wikipedia.org/wiki/Apollo_Guidance_Computer.
- [2] Moursund, David G. "Brief introduction to educational implications of Artificial Intelligence." (2006).
- [3] De Koster, René, Tho Le-Duc, and Kees Jan Roodbergen. "Design and control of warehouse order picking: A literature review." *European journal of operational research* 182.2 (2007): 481-501.
- [4] Frazelle, Edward, and Ed Frazelle. *World-class warehousing and material handling*. Vol. 1. New York: McGraw-Hill, 2002.
- [5] Guo, Anhong, et al. "Order picking with head-up displays." *Computer* 48.6 (2015): 16-24.
- [6] "Imagination in Action." Performance by Thad Starner, *Symbiotic AI: Creating an AI with Human Level Experience | Thad Starner | MIT 2019*, 12 May 2019, www.youtube.com/watch?v=PrhFMcvRhes.

[7] Kotipalli, Pramod. "Symbiotic Artificial Intelligence." *Pramod Kotipalli*, 22 June 2019, p13i.io/posts/2019/06/symbiotic-ai/.

[8] Haynes, Malcolm, and Thad Starner. "Effects of lateral eye displacement on comfort while reading from a video display terminal." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (2018): 1-17.

[9] Bambach, Sven, et al. "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[10] Betancourt, Alejandro, et al. "A sequential classifier for hand detection in the framework of egocentric vision." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014.

[11] Erol, Ali, et al. "Vision-based hand pose estimation: A review." *Computer Vision and Image Understanding* 108.1-2 (2007): 52-73.

[12] Lugaresi, Camillo, et al. "MediaPipe: A Framework for Perceiving and Processing Reality." (2019).

[13] Rong, Weibin, et al. "An improved CANNY edge detection algorithm." *2014 IEEE International Conference on Mechatronics and Automation*. IEEE, 2014.

[14] El-Khamy, Said E., Mona Lotfy, and N. El-Yamany. "A modified fuzzy Sobel edge detector." *Proceedings of the Seventeenth National Radio Science Conference. 17th NRSC'2000 (IEEE Cat. No. 00EX396)*. IEEE, 2000.

[15] Li, Youguo, and Haiyan Wu. "A clustering method based on K-means algorithm." *Physics Procedia* 25 (2012): 1104-1109.

[16] Contributors, TechTarget. "What Is Data Labeling? - Definition from WhatIs.com." *WhatIs.com*, TechTarget, 31 Aug. 2019, [whatis.techtarget.com/definition/data-labeling](https://www.techtarget.com/definition/data-labeling).

[17] Merkuryev Y., Burinskiene A., Merkuryeva G. (2009) Warehouse Order Picking Process. In: Merkuryev Y., Merkuryeva G., Piera M., Guasch A. (eds) *Simulation-Based Case Studies in Logistics*. Springer, London.

https://doi.org/10.1007/978-1-84882-187-3_9