**Gesture Recognition for Enabling Control of Electrical Devices**

by David T. Tung

B.S. in Mechanical Engineering, National Sun Yat-Sen University

Master of Computer Science, University of Southern California

A Praxis submitted to

The Faculty of

The School of Engineering and Applied Science

of The George Washington University

in partial fulfillment of the requirements

for the degree of Doctor of Engineering

August xx, 2025

Praxis directed by

Shahryar Sarkani  
 Adjunct Professor of Engineering Management and Systems Engineering

John Fossaceca  
 Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that David T. Tung has passed the Final Examination for the degree of Doctor of Engineering as of August xx, 2025. This is the final and approved form of the Praxis.

**Gesture Recognition for Enabling Control of Electrical Devices**

David T. Tung

Praxis Research Committee:

Shahryar Sarkani, Adjunct Professor of Engineering Management and Systems Engineering, Praxis Co-Director

John Fossaceca, Professorial Lecturer of Engineering Management and Systems Engineering, Praxis Co-Director

-------, Associate Professor of Engineering and Applied Science, Committee Member [place holder]

© Copyright 2025 by David T. Tung

All rights reserved

# Dedication

I would like to dedicate this research to my family [todo]

# Acknowledgements

I wish to thank Dr. Shahryar Sarkani and Dr. John Fossaceca for their guidance and support throughout the studying, research, and writing processes. [todo]

# Abstract of Praxis

**Gesture Recognition for Enabling Control of Electrical Devices**

Individuals with mobility impairments, particularly wheelchair users, often face challenges in accessing and controlling electrical devices. A gesture-based control system can greatly enhance accessibility for this group. This research praxis aims to develop and evaluate a user-friendly, cost-effective machine-learning approach that enables users to control devices through pointing gestures. Specifically, it focuses on extending DeePoint, a 3D pointing direction prediction model (Nakamura et al., 2023), into a unified machine learning system capable of identifying the electrical devices a user is pointing at. The goal is to lay the groundwork for future applications that empower individuals with mobility impairments to intuitively and efficiently control household devices using simple, natural gestures.

**Table of Contents**

[Dedication iv](#_Toc179138418)

[Acknowledgments v](#_Toc179138419)

[Abstract of Praxis vi](#_Toc179138420)

[List of Figures ix](#_Toc179138421)

[List of Tables x](#_Toc179138422)

[List of Symbols xi](#_Toc179138423)

[List of Acronyms xii](#_Toc179138424)

[Chapter 1—Introduction 1](#_Toc179138425)

[1.1 Background 1](#_Toc179138426)

[1.2 Research Motivation 3](#_Toc179138427)

[1.4 Problem Statement 4](#_Toc179138428)

[1.4 Thesis Statement 4](#_Toc179138429)

[1.5 Research Objectives 4](#_Toc179138430)

[1.6 Research Questions and Hypotheses 6](#_Toc179138431)

[1.7 Scope of Research 6](#_Toc179138432)

[1.8 Research Limitations 7](#_Toc179138433)

[1.9 Organization of Praxis 7](#_Toc179138434)

[Chapter 2—Literature Review 9](#_Toc179138435)

[2.1 Introduction 9](#_Toc179138436)

[2.2 Another Section 9](#_Toc179138437)

[2.3 Another Section 12](#_Toc179138438)

[Chapter 3—Methodology 19](#_Toc179138439)

[3.1 Introduction 19](#_Toc179138440)

[3.2 Another Section 19](#_Toc179138441)

[Chapter 4—Results 20](#_Toc179138442)

[4.1 Introduction 20](#_Toc179138443)

[4.2 Another Section 20](#_Toc179138444)

[Chapter 5—Discussion and Conclusions 21](#_Toc179138445)

[5.1 Discussion 21](#_Toc179138446)

[5.2 Conclusions 21](#_Toc179138447)

[5.3 Contributions to Body of Knowledge 21](#_Toc179138448)

[5.4 Recommendations for Future Research 21](#_Toc179138449)

[References 22](#_Toc179138450)

[Appendix A 30](#_Toc179138451)

# List of Figures[todo]

[Figure 1. The architecture of the device classification system. 3](#_Toc179180953)

[Figure 4-1. XYZ. **Error! Bookmark not defined.**](#_Toc179180954)

[Figure A-1. Histogram of XYZ. **Error! Bookmark not defined.**](#_Toc179180955)

# List of Tables[todo]

[Table 4-1. Pearson Correlations Between W and T **Error! Bookmark not defined.**](#_Toc519272309)

[Table A-1. Parametric Correlations of X and Y **Error! Bookmark not defined.**](#_Toc519272310)

# List of Symbols

State of the system

Output of the system

Noise

# List of Acronyms

ADA Americans with Disabilities Act

YOLO You Only Look Once is a real-time object detection system

CNN Convolutional Neural Networks

Fast R-CNN Fast Region-Based Convolutional Neural Networks

DL Deep Learning

HRI Human-Robot Interaction

HOI Human Object Interaction

ViT Vision Transformer

CLIP Contrastive Language-Image Pretraining

NLP Natual Language Processing

ML Machine Learning

CMU

U.S. United States

# Chapter 1—Introduction

## 1.1 Background

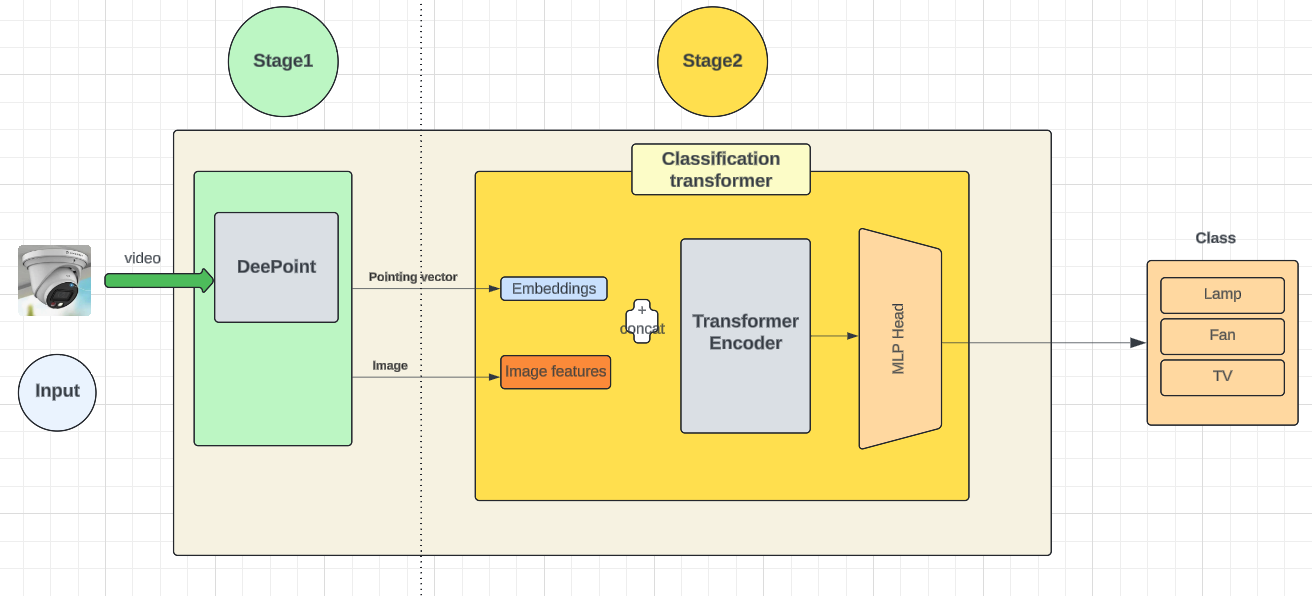
As the number of older adult households continues to grow in the decades ahead, more homeowners are expected to undertake expensive accessibility renovations. One of these projects' most critical and costly aspects is improving access to electrical devices, particularly for individuals with mobility impairments who rely on wheelchairs for movement. Accessible housing design features can significantly enhance the independence of people with mobility disabilities within their homes. In addition, the number of caregivers is shrinking as the need for care explodes (Nora S. *et al*., 2020). Despite advances in assistive devices, laws, and technology aimed at improving safety and aiding daily activities, people with mobility disabilities continue to face barriers in accessing electrical devices, reducing their capacity to live independently.

Over the years, various solutions have been developed to assist individuals with mobility impairments to control household devices, including voice-activated systems and augmented reality or mixed reality smart glasses (Zhou K. *et al*., 2023). Despite their potential, these technologies are not yet ready to be widespread within this group due to several challenges. Voice-activated systems, for instance, often underperform in noisy environments, making them unreliable in busy or public settings. Additionally, they are unsuitable for environments where quiet is essential, such as a baby’s room or a shared living space. While smart glasses provide a hands-free solution, they come with drawbacks, such as high costs and discomfort when worn for extended periods, which limits their practicality for long-term, daily use.

Moreover, many existing gesture-based control systems rely on specialized hardware, like RGB-D sensors, which can be effective but are often expensive, bulky, and cumbersome to use and maintain. These limitations prevent widespread adoption, particularly for individuals seeking affordable and convenient solutions to improve their independence. As a result, there is a growing need for more user-friendly, intuitive, and cost-effective approaches to overcome these barriers and provide greater accessibility in daily device control (Chang *et al*. 2022).

Our solution uses finger-pointing gestures, a natural and efficient way to indicate objects or devices of interest. Implementing an intuitive system that allows device control through pointing gestures could significantly benefit individuals with mobility impairments. For example, a person in a wheelchair could navigate a room and point at devices like lights or appliances to control them, with the scene being captured by cameras mounted in the corners, which may already be available or simple to install.

In this work, illustrated in Fig. 1, We propose a multi-stage machine learning system that extends the 3D pointing direction prediction model, DeePoint (Nakamura et al., 2023). This system incorporates a device classification transformer, which utilizes the predicted pointing direction and images from DeePoint's outputs to form a unified device classification system capable of identifying which devices the user is pointing at. This classification system can process video input that monitors the wheelchair user, detecting their intent to control devices by pointing at the device and generating a command for the device to act[[1]](#footnote-2).



##### Figure 1. The architecture of the device classification system.

With the proposed classification system, our contribution establishes a foundation for future touchless device control applications extending beyond household devices to include elements like elevator buttons. This system aims to empower individuals with mobility impairments to intuitively and efficiently control devices through natural pointing gestures

## 1.2 Research Motivation

The motivation behind this research arises from the persistent challenges faced by individuals with mobility impairments, particularly wheelchair users, in controlling household devices. While offering a degree of autonomy, existing solutions like voice-activated systems and smart glasses have notable limitations, such as reduced effectiveness in noisy environments, physical discomfort, and high costs. Additionally, hardware-dependent systems, such as those utilizing RGB-D sensors, are often expensive and cumbersome, further restricting accessibility.

We need more natural, intuitive, and affordable solutions that enable individuals with mobility impairments to interact seamlessly with devices. Gesture recognition, especially through pointing direction, offers a promising alternative. However, research on the visual interpretation of 3D pointing gestures is still limited (Nakamura et al., 2023). To bridge this gap, this study seeks to develop a machine-learning model that leverages standard RGB cameras for 3D pointing recognition to provide a practical and cost-effective solution for device control.

## 1.4 Problem Statement

## *Individuals using wheelchairs often face challenges accessing electrical devices, so approximately 35% of U.S. housing units may need to be modified to meet the accessibility requirements of the devices (U.S. Department of Housing and Urban Development, n.d., 2015).*

## Even in places that comply with the Americans with Disabilities Act (ADA), individuals using wheelchairs are often required to stretch to reach electrical switches.

## Furthermore, Assistive devices, such as canes, robotic arms, intelligent eyewear, mobile phones, or AI-enabled wearables, are expensive.

## 1.4 Thesis Statement

*A two-stage classification system is needed to identify the electrical devices a wheelchair user points at, enabling touchless device control and enhancing accessibility.*

## The primary research product is a device classification system developed in Python, designed to enhance device accessibility for wheelchair users. This technology enables integrators and developers to improve accessibility. The research presents a new contribution by developing a classification system that identifies electrical devices pointed at by a wheelchair user, leveraging DeePoint and object detection as its core machine learning technologies. The methodology incorporates machine learning, transformers, deep learning, computer vision, and object detection. Input data will consist of videos of a seated individual pointing at electrical devices, while the system’s output will classify the device and provide a probability score.

## 1.5 Research Objectives

The primary objective of this research is to develop a machine-learning model that can classify devices a wheelchair user is pointing at. The key research focuses on exploring and evaluating the classification system’s accuracy and performance across various network architectures and scenarios. Detailed objectives are as follows:

**RO1**. Evaluate the feasibility of a machine learning device classification system that integrates pointing direction prediction with object detection capabilities. Assess the system’s performance using test data to explore its effectiveness and accuracy.

**RO2**. Investigate the impact of using different pointing direction prediction models as components within the proposed classification system, comparing their accuracy and efficiency, specifically DeePoint vs. OpenPose.

**RO3**. Analyze the impact of using different object detection models as components within the proposed pointing device classification system, comparing their accuracy and efficiency, specifically YOLO vs. Fast R-CNN.

## 1.6 Research Questions and Hypotheses

This study aims to clarify and explain the following three research questions:

**RQ1:** Does tracking gaze direction enhance the accuracy of pointing direction prediction in the first stages of the two-stage classification system?

**RQ2:** Can a two-stage classification system be developed to identify the electrical devices a wheelchair user points at, enabling touchless device control and improving accessibility?

**RQ3:** Which ML model works best for the second stage of the two-stage classification system for identification of electrical devices, as pointed by wheelchair user?

**H1:** Tracking gaze direction can improve the accuracy of pointing direction prediction by approximately 5% in the first stages of a two-stage classification system.

**H2:** The proposed two-stage classification system can reach 70% accuracy in identifying the electrical devices a wheelchair user points at.

**H3:** In the second stage, the proposed device classification system incorporating the YOLO is expected to outperform the model using the Fast R-CNN as an object detection component.

## 1.7 Scope of Research

The scope of this praxis is to evaluate the feasibility of developing a device classification system capable of interpreting gesture-based interactions to improve device accessibility for wheelchair users. The proposed solution extends a pointing direction prediction model and integrates a device classification transformer. Furthermore, the actual electrical device control part is not in the scope.

## 1.8 Research Limitations

The following factors limit this research:

While incorporating a confirmation step for pointing gestures would greatly improve reliability and user experience in practical applications, it falls outside the scope of this study. Instead, the focus is on developing a device classification model.

This research intentionally limits the device categories to three common household items: a TV, a Fan, and a Lamp—to streamline model development and evaluation. Future work could expand this range to provide a more comprehensive solution for individuals with mobility impairments.

Additionally, the experiments will be conducted exclusively in indoor environments, and the test data is derived from videos of seated users rather than actual wheelchair users.

## 1.9 Organization of Praxis

This Praxis consists of five chapters, as follows:

Chapter 1 begins with the background and the research motivation, then continues with the research objectives, questions, and hypotheses. It ends with the scope of the research and the limitations of this research.

Chapter 2 presents a review of relevant literature, beginning with the DeepPoint paper (Nakamura *et al*., 2023), which serves as the foundation for this praxis. Additional literature provides context on accessibility challenges faced by wheelchair users (JCHS, Harvard, 2023; U.S. Department of Housing and Urban Development, n.d., 2015). The WorldPoint paper from CMU offers technical insights into the implementation of the ray-casting algorithm for object intersection (Kim D. et al., 2023). Furthermore, the MultiNet framework demonstrates how multiple models can be effectively combined into one (Teichmann M. et al., 2018). The end-to-end human-object interaction detection paper presents a solution based on the HOI pattern. (Zou, C. *et al*., 2018).

Chapter 3 presents the three research questions used in this praxis and testing of the hypotheses.

Chapter 4 covers the results and analyses of the statistical methods presented in Chapter 3.

Chapter 5 closes the praxis with a discussion of results and a conclusion. It also includes discussions of the contributions to the body of knowledge and recommendations for future research in the area.

# Chapter 2—Literature Review

## 2.1 Introduction

The increasing prevalence of mobility impairments has underscored the need for effective assistive technologies that enable users to control devices with minimal physical effort. Among all of the assistive technologies, gestures have emerged as a natural and intuitive method for human-device interaction, allowing users to control devices through simple gestures (Islam, M.M., 2020). This chapter provides a comprehensive literature review of the topics related to assistive technologies, finger-pointing gesture recognition, and transformer and neural network algorithms, which enable the core of intelligent assistive technology for mobility-impaired users. The purpose of this review is to provide a summary of the research that has been published on these topics and to analyze the existing body of technical knowledge.

The chapter begins by reviewing mobility-impaired users' challenges and the need for assistive technologies. It then continues by analyzing the literature regarding existing solutions for these users. This chapter will also give a review of current research on gesture recognition, object detection, and human-object interaction.

Next, the chapter reviews current research on AI and neural networks, especially the Transformer and Convolution Neural Network (CNN) approach used in this practice. The chapter concludes with a summary of the findings and potential implications of the current study.

## 2.2 Assistive living and technologies review

America's aging population is undergoing unprecedented growth, and a significant portion faces mobility challenges, with many requiring wheelchairs to move around and access daily utilities. As individuals age, they are increasingly likely to experience mobility disabilities, which pose serious challenges to independent living. Many homes in the US are not equipped for such needs, often requiring costly modifications to meet accessibility standards. The financial burden of these adjustments is significant. In addition to modification, according to the US Department of Health and Human Services (HHS), nearly 70 percent of people who reach the age of 65 will require some form of long-term care in their lifetime. This additional care and housing modifications can be overwhelming, particularly for those already facing financial constraints.

Compounding this issue is the shrinking number of available caregivers at a time when the demand for long-term care is surging (Nora S. *et al*., 2020). With fewer caregivers to provide assistance, new solutions are urgently needed to bridge the gap. Technology has the potential to play a transformative role in addressing these challenges. Innovations in assistive technology can significantly improve the quality of life for older adults by providing them with the tools to regain a level of independence.

For instance, Chen, W. L. et al. introduced a novel home appliance control system tailored for individuals with disabilities. This system enables them to perform daily tasks independently. Such systems represent a step toward greater autonomy for people with mobility issues, allowing them to control household devices with minimal physical effort. Another example is the work of Bourbakis, N.G., who proposed an intelligent system that integrates robots, sensors, and other assistive technologies to aid with mobility. While this system provides a comprehensive solution for those with severe mobility impairments, it is also prohibitively expensive for widespread adoption. The combination of advanced robotics, artificial intelligence, and sensor technologies presents an impressive solution, but it may only be feasible for those with considerable financial resources or specialized needs.

However, as with many technological solutions, these advances come with their own set of challenges. The primary hurdles involve the cost of these systems and the obtrusiveness of the devices. High upfront costs can put these technologies out of reach for many older adults, particularly those on fixed incomes. Furthermore, the physical presence of devices in the home can be intrusive, potentially disrupting the comfort and aesthetics of the living environment. For a practical application in solving accessibility issues for older adults, the challenge lies in developing affordable and unobtrusive technologies while still being effective. Solutions must focus on functionality and user experience, ensuring that devices blend seamlessly into the home environment without being overwhelming or difficult to use. Affordability is key, especially as the population ages and the number of individuals needing assistance continues to rise.

In response to these challenges, intuitive, low-cost solutions such as gesture recognition systems or simple control interfaces could be designed to allow individuals to control devices through natural interactions with AI technology and without the need for complex hardware or invasive modifications (Islam, M.M., 2020). These systems could provide a cost-effective alternative, allowing older adults to easily manage their daily tasks while avoiding the high costs associated with robotics and sensor-heavy systems. By focusing on accessible, affordable, and unobtrusive technological solutions, we can help bridge the gap between the increasing need for care and the dwindling number of caregivers. These technologies have the potential to empower older adults, enabling them to live more independently and with greater dignity as they age. (Courtney, K. L. et al, 2007; Moon NW, et al, 2019)

## 2.3 Pointing gestures recognition and object interaction

Gesture control has been widely adopted in the AR/VR industry. The egocentric vision, also known as first-person vision, usually refers to capturing and processing images and videos from cameras worn on a person’s head. With the development of smart wearable cameras and augmented reality headsets such as Meta Oculus, Microsoft HoloLens, and Google Glass, egocentric vision and its potential applications have drawn much attention. This 2016 CVPR paper, “A Pointing Gesture-Based Egocentric Interaction System: Dataset, Approach, and Application” (Huang, Y., Liu, et al., 2016), researches AR-based pointing technology, especially hand gesture-based interaction. This paper presents a solution for point gesture-based interaction in egocentric vision and its applications. Firstly, a dataset named EgoFinger is established, focusing on pointing gestures for egocentric vision. Furthermore, they propose a two-stage Faster R-CNN-based hand detection and dual-target fingertip detection framework. Later, Cao, C et al,. proposed an egocentric gesture recognition using recurrent CNN with spatiotemporal transformer modules for wearable AR device movement problem. Alam M. M et al., introduce a unified learning approach to predict both the probabilistic output of the egocentric gesture of fingers and the positional output of all the fingertips using one forward propagation of a CNN. For special hardware, G. Park et al develops a gesture recognition method with special hardware ( radar and attena) and deep learning model.

The wearable solution seems to be an expensive and obstructive option for mobility-impaired users. Nakamura et al., 2023, realize automatic visual recognition and direction estimation of pointing for a non-wearable gesture recognition solution. This paper introduces the first neural pointing understanding method and the first-of-its-kind large-scale dataset for pointing recognition and direction estimation; this dataset consists of more than 2 million frames of 33 people pointing in various styles, annotated for each frame with pointing timings and 3D directions. Through extensive experiments, the accuracy and efficiency of DeePoint are demonstrated, and it is believed that the DP Dataset and DeePoint can serve as a sound foundation for visual human intention understanding.

WorldPoint (Kim D. et al., 2023) is an innovative idea that uses pointing gestures as a rapid and natural trigger for mobile interactions. This method does not require wearable hardware, but users must always carry a mobile phone. It is the work of Kim D. et al. at CMU, who utilize the recent inclusion of wide-angle, rear-facing smartphone cameras and hardware-accelerated machine learning to enable real-time, infrastructure-free, finger-pointing interactions on today’s mobile phones.

Recently, pointing gesture technologies have become popular in the automobile industry. In the paper *"You Have a Point There: Object Selection Inside an Automobile Using Gaze, Head Pose, and Finger Pointing"* (Aftab *et al*. 2020), finger-pointing technology is explored for automotive user interaction. The automotive industry is rapidly advancing in user interaction technologies, with mid-air gestures and voice commands already enhancing driver-vehicle interaction (Fig. 2). This paper proposes a multimodal fusion method: gaze, head pose, and finger-pointing gestures, using speech solely as a trigger for the fusion process. This paper compared state-of-the-art deep neural network architectures with traditional machine learning; the results indicate that deep learning methods significantly improve pointing direction accuracy when integrating multiple modalities. This multimodal approach has the potential to enhance user interaction in vehicles, laying the foundation for future applications that rely on sensor fusion for a more intuitive and responsive driving experience. In our work, we utilized a software-based approach (DeepPoint) to replace the need for expensive gesture camera hardware, significantly reducing the overall cost. The problem of object selection inside a car has also been presented by Roider et al., who integrate eye gaze with finger-pointing gestures in a passive manner using a simple rule-based fusion approach. They have shown that the selection on an in-vehicle display screen achieves increased pointing accuracy over a single modality, i.e., finger pointing (Roider et al., 2018). This experiment is limited to only four objects on a screen adjacent to each other.

A finger pointing at a car dashboard

Description automatically generated

##### Figure 2. Driver makes a pointing gesture to interact with the car. ( Source: Aftab, A. R., et al., 2020)

In the robotics industry, gestures are a common way for human-robot interaction; the 2023 ECCV paper *"Interactive Multimodal Robot Dialog Using Pointing Gesture Recognition"* (Tanada, K et al., 2024), pointing gestures are identified as an intuitive form of human-robot interaction. This work proposes a system for interactive, multimodal, task-oriented robot dialog that leverages pointing gesture recognition. The system integrates state-of-the-art computer vision techniques to recognize objects, hand positions, orientations, and overall human poses, allowing for a comprehensive understanding of pointing gestures and the corresponding target objects. Furthermore, M. Ürkmez and H. I. Bozma proposed a two-stage CNN approach to detect 3D hand-pointing direction. However, this method requires a depth camera and focus on HRI ( Human-Robot interaction). Similar to robots, pointing gestures are also used in drone applications; for the paper titled: Using Pointing Gesture to Define a Target Object (Medeiros et al. 2020), which develops a method for firefighters to specify the place got fire and direct drone to fly over to the location. Another similar paper by Medeiros et al. integrates depth info with SLAM to achieve the same purpose.

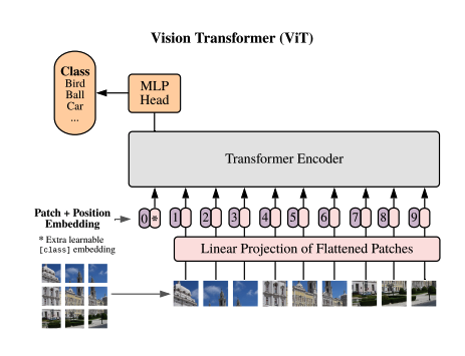
Human-object interaction (HOI) detection plays an important role in the high-level human-centric scene understanding and has attracted considerable research interest recently. The goal of HOI detection is to localize humans and objects, as well as recognize the interaction between them. Previous studies (Chen Gao. 2018.; Chen Gao. 2020) present promising results on HOI detection by decoupling this task into object detection and interaction classification (a two-stage approach). More recent approaches (Tiancai Wang; Kim B.; Yue L.) have introduced a surrogate interaction detection problem to optimize HOI detection indirectly (one-stage approach). Finally, Zou, C et al., proposed an end-to-end HOI prediction directly without the need for a multiple-stage approach, which can possibly be applied to gesture-object interaction.

**2.4 Vision transformer technology**

This section provides an overview of Vision Transformer technology and references the key papers related to its development.

The core of Transformer technology is the attention mechanism. The attention mechanism is a concept used in machine learning models to help them focus on specific parts of input data rather than processing all information equally. It allows the model to "attend" to the most relevant pieces of data when making predictions. In sequence models like Transformers, attention calculates the importance of each word or token in relation to others, enabling the model to capture long-range dependencies and contextual relationships more effectively. In other words, an attention function can be understood as a process that takes a query and a set of key-value pairs as inputs and generates an output. Each of these (query, keys, values, output) is represented as a vector. The output is a weighted sum of the values, where the weight for each value is determined by how closely the query matches the corresponding key, computed through a compatibility function. This allows the model to focus on relevant information in the data. (Vaswani et al., 2017). The result is to make attention highly useful in natural language processing and image recognition tasks)

The attention-based Transformer architecture has become the de-facto standard for natural language processing tasks, but its applications to computer vision remain limited. In vision, CNNs remain dominant in computer vision ( LeCun et al., 1989; Krizhevsky et al., 2012); focus is either applied in conjunction with convolutional networks or used to replace certain components of convolutional networks while keeping their overall structure in place. Dosovitskiy et al. at Google Brain published a paper titled: An image is worth 16x16 words: Transformers for image recognition at Scale (Dosovitskiy et al., 2020), which changed the NLP-only usage in transformers by introducing Vission Transformer (see Figure 3). This paper explores directly applying them to images by treating image patches as tokens for classification. In our proposal architecture (see Figure 1), we have integrated transformer technology by leveraging a Vision Transformer (ViT) architecture.



##### Figure 3. Vision Transformer Architecture ( Source: Dosovitskiy et al., 2020)

Architectures for video understanding have mirrored advances in transformer-based image recognition. In the Video Vision Transformer paper (Arnab et al., 2021), they develop pure-transformer architectures for video classification inspired by ViT and the fact that attention-based architectures are intuitive for modeling long-range contextual relationships in video.

To pair text with image in transformer, CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) is a model developed by OpenAI that bridges the natural language and visual understanding. It is trained on a large dataset of text-image pairs and uses contrastive learning to align images with their corresponding text descriptions. CLIP can perform tasks such as zero-shot image classification, where it matches images to appropriate labels without needing task-specific fine-tuning. It represents a significant step in multimodal AI by enabling flexible and powerful connections between language and visual content, making it versatile across many applications. Due to CLIP’s success, there are many variations of CLIP, such as Zeng et al. present CLIP2, which directly learns the transferable 3D point cloud representation in realistic scenarios with a novel proxy alignment mechanism.

**2.5 Object Detection Technology**

This section provides an overview of object detection technology and references the key papers related to its development.

**2.6 Summary and Conclusion**

This literature review section provided a comprehensive examination of journal papers, conference papers, and books on product defect management, the usage of customer text reviews by companies and engineering teams, and the tools for analyzing textual data with a focus on RNN and LDA. Based on the literature review, manufacturing companies are still facing huge product recall problems even with modern quality assurance technologies and tools, and online customer opinions often get overlooked. Although the literature review has identified a few methods of using OCRs to identify product defect information, these methods are either highly domain-specific, accepting only certain negative OCRs, or identifying defect information only at the OCRs level. This study introduces the newly mature RNN and LDA method to provide solutions for identifying defective product insights and bridging the knowledge gap between product defect management, customer feedback, and neural networking.

# Chapter 3—Methodology

## 3.1 Introduction

## 3.2 Another Section

# Chapter 4—Results

## 4.1 Introduction

## 4.2 Another Section

# Chapter 5—Discussion and Conclusions

## 5.1 Discussion

## 5.2 Conclusions

## 5.3 Contributions to Body of Knowledge

1. [todo]

## 5.4 Recommendations for Future Research

# References

U.S. Department of Housing and Urban Development. (n.d.). Accessibility of America’s housing stock: Analysis of the 2011 American Housing Survey. Office of Policy Development and Research.

Bureau of Labor Statistics, U.S. Department of Labor. (n.d.). *Occupational Outlook Handbook*. Retrieved from <https://www.bls.gov/ooh/>

U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. (2022). *What is the lifetime risk of needing and receiving long-term services and supports?* <https://aspe.hhs.gov/reports/what-lifetime-risk-needing-receiving-long-term-services-supports-0>

Nora Super, Three Trends Shaping the Politics of Aging in America, *Public Policy & Aging Report*, Volume 30, Issue 2, 2020, Pages 39-45, <https://doi.org/10.1093/ppar/praa006>

Bourbakis, N.G. (2022). Challenges in Assistive Living Based on Tech Synergies: The Cooperation of a Wheelchair and A Wearable Device. In: Tsihrintzis, G.A., Virvou, M., Esposito, A., Jain, L.C. (eds) Advances in Assistive Technologies. Learning and Analytics in Intelligent Systems, vol 28. Springer, Cham. <https://doi.org/10.1007/978-3-030-87132-1_11>

Chen, W. L., Liou, A. H. A., Chen, S. C., Chung, C. M., Chen, Y. L., & Shih, Y. Y. (2007). A novel home appliance control system for people with disabilities. *Disability and Rehabilitation: Assistive Technology*, *2*(4), 201–206. <https://doi.org/10.1080/17483100701456012>

Islam, M.M., Islam, M.R. & Islam, M.S. An Efficient Human Computer Interaction through Hand Gesture Using Deep Convolutional Neural Network. *SN COMPUT. SCI.* **1**, 211 (2020). https://doi.org/10.1007/s42979-020-00223-x

Courtney, K. L., Demiris, G., & Hensel, B. K. (2007). Obtrusiveness of information-based assistive technologies as perceived by older adults in residential care facilities: A secondary analysis. *Medical Informatics and the Internet in Medicine*, *32*(3), 241–249. <https://doi.org/10.1080/14639230701447735>

Moon NW, Baker PM, Goughnour K. Designing wearable technologies for users with disabilities: Accessibility, usability, and connectivity factors. Journal of Rehabilitation and Assistive Technologies Engineering. 2019;6. doi:10.1177/2055668319862137

Jirak, D., Biertimpel, D., Kerzel, M. *et al.* Solving visual object ambiguities when pointing: an unsupervised learning approach. *Neural Comput & Applic* **33**, 2297–2319 (2021). https://doi.org/10.1007/s00521-020-05109-w

Nakamura, S., Kawanishi, Y., Nobuhara, S., & Nishino, K. (2023). DeePoint: Visual Pointing Recognition and Direction Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 20577-20587).

Huang, Y., Liu, X., Zhang, X., & Jin, L. (2016). A pointing gesture based egocentric interaction system: Dataset, approach and application. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 16-23).

Cao, C., Zhang, Y., Wu, Y., Lu, H., & Cheng, J. (2017). Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE international conference on computer vision* (pp. 3763-3771).

Alam, M. M., Islam, M. T., & Rahman, S. M. (2022). Unified learning approach for egocentric hand gesture recognition and fingertip detection. *Pattern recognition*, *121*, 108200.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 482–490, 2020.

Kim Bumsoo, Choi Taeho, Kang Jaewoo, and J. Kim Hyunwoo. Uniondet: Union-level detector towards real-time human-object interaction detection. In European Conference on Computer Vision. Springer, 2020.

Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4116–4125, 2020.

Gao, C., Xu, J., Zou, Y., & Huang, J. B. (2020). Drg: Dual relation graph for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16* (pp. 696-712). Springer International Publishing.Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instancecentric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437, 2018.

M. Ürkmez and H. I. Bozma, "Detecting 3D Hand Pointing Direction from RGB-D Data in Wide-Ranging HRI Scenarios," *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Sapporo, Japan, 2022, pp. 441-450, doi: 10.1109/HRI53351.2022.9889385.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16901-16911)

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440-1448).

Constantin, S., Eyiokur, F. I., Yaman, D., Bärmann, L., & Waibel, A. (2023). Multimodal Error Correction with Natural Language and Pointing Gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1976-1986).

Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., ... & Sun, J. (2021). End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11825-11834).

Kim, D., Mollyn, V., & Harrison, C. (**2023**). WorldPoint: Finger Pointing as a Rapid and Natural Trigger for In-the-Wild Mobile Interactions. *Proceedings of the ACM on Human-Computer Interaction*, *7*(ISS), 357-375.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836-6846).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541-551

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

G. Park, V. K. Chandrasegar and J. Koh, "Hand Gesture Recognition using Deep learning Method," *2021 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (APS/URSI)*, Singapore, Singapore, 2021, pp. 1347-1348, doi: 10.1109/APS/URSI47566.2021.9703901.

Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., ... & Xu, H. (2023). CLIP2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15244-15253). Zhou, Kanglei & Chen, Chen & Ma, Yue & Leng, Zhiying & Shum, Hubert & Li, Fred & Liang, Xiaohui. (2023). A Mixed Reality Training System for Hand-Object Interaction in Simulated Microgravity Environments. 167-176. 10.1109/ISMAR59233.2023.00031.

Medeiros, A.C.S., Ratsamee, P., Uranishi, Y., Mashita, T., Takemura, H. (2020). Human-Drone Interaction: Using Pointing Gesture to Define a Target Object. In: Kurosu, M. (eds) Human-Computer Interaction. Multimodal and Natural Interaction. HCII 2020. Lecture Notes in Computer Science(), vol 12182. Springer, Cham. <https://doi.org/10.1007/978-3-030-49062-1_48>

Medeiros, A.C.S., Ratsamee, P., Orlosky, J. *et al.* 3D pointing gestures as target selection tools: guiding monocular UAVs during window selection in an outdoor environment. *Robomech J* **8**, 14 (2021). https://doi.org/10.1186/s40648-021-00200-w

Xu, J., Wang, H., Zhang, J., & Cai, L. (2022). Robust hand gesture recognition based on RGB-D Data for natural human–computer interaction. *IEEE Access*, *10*, 54549-54562.

Chang, V., Eniola, R. O., Golightly, L., & Xu, Q. A. (2023). An Exploration into Human–Computer Interaction: Hand Gesture Recognition Management in a Challenging Environment. *SN Computer Science*, *4*(5), 441.

Pelgrim, M. H., He, I. X., Lee, K., Pabari, F., Tellex, S., Nguyen, T., & Buchsbaum, D. (2024). Find it like a dog: Using Gesture to Improve Object Search. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Jirak, D., Biertimpel, D., Kerzel, M., & Wermter, S. (2021). Solving visual object ambiguities when pointing: an unsupervised learning approach. *Neural Computing and Applications*, *33*, 2297-2319.

Etoh, M., Tomono, A., & Kobayashi, Y. (1989). Direct Finger Pointing as a Man-machine Interface. *IFAC Proceedings Volumes*, *22*(12), 125-130.

Lee, K. (2024). *Point it Out: Using Gesture to Improve Object Search* (Doctoral dissertation, Brown University Providence, Rhode Island).

Wong, N., & Gutwin, C. (2010, April). Where are you pointing? The accuracy of deictic pointing in CVEs. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1029-1038).

Rümelin, S., Marouane, C., & Butz, A. (2013, October). Free-hand pointing for identification and interaction with distant objects. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 40-47).

Aftab, A. R., von der Beeck, M., & Feld, M. (2020, October). You have a point there: object selection inside an automobile using gaze, head pose and finger pointing. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 595-603).

Florian Roider and Tom Gross. 2018. I See Your Point: Integrating Gaze to Enhance Pointing Gesture Accuracy While Driving. In Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. 351–358.

Harika, M. (2016). *Finger-pointing gesture analysis for slide presentation* (Doctoral dissertation, 부경대학교 대학원).

Erden, F., & Cetin, A. E. (2014). Hand gesture based remote control system using infrared sensors and a camera. *IEEE Transactions on Consumer Electronics*, *60*(4), 675-680.

Erden, F., Velipasalar, S., Alkar, A. Z., & Cetin, A. E. (2016). Sensors in assisted living: A survey of signal and image processing methods. *IEEE Signal Processing Magazine*, *33*(2), 36-44.

Sudhakar, S., Liu, R., Van Hoorick, B., Vondrick, C., & Zemel, R. (2024). Controlling the World by Sleight of Hand. *arXiv preprint arXiv:2408.07147*.

Nelson A, McCombe Waller S, Robucci R, Patel C, Banerjee N. Evaluating touchless capacitive gesture recognition as an assistive device for upper extremity mobility impairment. Journal of Rehabilitation and Assistive Technologies Engineering. 2018;5. doi:10.1177/2055668318762063

Tanada, K., Matsuzaki, S., Tanaka, K., Nakaoka, S., Kondo, Y., & Mori, Y. Pointing Gesture Understanding via Visual Prompting and Visual Question Answering for Interactive Robot Navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.

# Appendix A

1. The actual device control research and implementation are not in the scope of this praxis. [↑](#footnote-ref-2)