Towards Natural Underwater Human-Robot Interaction: Pointing Gesture Recognition for Autonomous Underwater Vehicles

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Andrea Maree Walker

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

ADVISOR

Junaed Sattar

May, 2021

Acknowledgements

I would first like to thank my research advisor Junaed Sattar for his guidance and support throughout my research. Without his advice, leadership, and mentorship, this work would not be possible. I would also like to thank the members of the Interactive Robotics and Vision lab for their support and collaboration - in particular Luoyao Chen for collaborating with me in studying pointing gestures, Karin de Langis and Michael Fulton for sharing their knowledge of HRI, and Sadman Sakib Enan and Md Jahidul Islam for their insight into understanding deep learning.

I would also like to thank Resha Tejpaul for her support and assistance in developing the Institutional Review Board documentation.

Additionally, I would like to thank my committee members Professor Ju Sun and Professor Natalia Perkins.

Finally, I would like to thank my parents for their ongoing love and support.

Dedication

To my parents, who have lovingly supported me from the beginning, giving me the best foundation for life possible. Thank you for teaching me since day one, always encouraging my curiosity and eagerness to learn, and being my ongoing support throughout my entire educational journey.

Abstract

Underwater robotics is a motivating field of research with a wide variety of both industrial and scientific applications. In particular, the development of autonomous underwater vehicles to assist divers in performing difficult, dangerous, or undesirable tasks has the potential to expand our abilities in the aquatic domain while reducing the risks presented to divers. For a diver and an autonomous underwater vehicle to work in collaboration, there must be an established interaction protocol; the study of such protocols is central to the field of human-robot interaction. In the underwater domain, the attenuation of both electromagnetic signals and sound limits these traditional communication protocols, leaving machine vision as the primary perception methodology. Thus gestures become a natural choice for diver-robot communication. Since pointing gestures are represented and recognized in cultures around the world, they serve as a foundational, natural gesture for divers in a demanding aquatic environment. Thus, in this work we lay the groundwork for implementing a pointing gesture recognition algorithm for use onboard autonomous underwater vehicles. Specifically, we contribute a human study of individuals performing four classes of pointing gestures, three datasets developed to study pointing gestures, and an analysis of four state-of-the-art object detection frameworks for recognizing pointing gestures in the aquatic domain.

Contents

Acknowledgements **i**

Dedication **ii**

Abstract **iii**

List of Tables **vi**

List of Figures **vii**

List of Abbreviations **ix**

1 Introduction **1**

1.1 Motivation for Underwater Robotics 2

1.2 Human-Robot Interaction for Autonomous Vehicles 3

1.3 Gestural-Based Communication 5

1.4 Pointing Gestures 6

1.5 Contributions 7

2 Related Works **8**

2.1 Human-Robot Interaction 8

2.2 Gestural-Based HRI 8

2.3 Gestural Communication in Underwater Robotics 11

3 Methodology **15**

3.1 Human Study 16

3.2 Dataset Preparation 20
3.3 Deep Learning Algorithms for Pointing Gesture Recognition 28

4 Experiments 33

4.1 Pointing Gesture Recognition Experiments 33

5 Conclusion and Future Work 44

Bibliography 46

Appendix A. IRB Human Study Details 56

A.1 Social Protocol, HRP-580 57
A.2 Social-Behavioral Consent Form, HRP-582 77
A.3 Exemption Determination 83
A.4 Information Sheet for Research, HRP-587 85
A.5 Final Approval Document 88
A.6 Promotional Materials 91
A.7 Intake Form 98

List of Tables

3.1 General Statistics on the Study Data Collected. 20

3.2 Dataset statistics for Data Split 1. 28

3.3 Dataset statistics for Data Split 2. 28

4.1 Model Feature Extraction Backbones and Baseline Training Checkpoints. 33

4.2 Inference speed on NVIDIA Titan Xp GPU. 37

4.3 Summary of AP and IOU Metrics across the two data splits and four object detectors. 40

4.4 Missed detections summary by dataset and object detector. 41

List of Figures

1.1 Underwater human-robot interaction scenarios with the (a) AQUA robot [1] and (b) LoCO AUV [2]. 2

1.2 Examples of the challenging, visually degraded imagery faced underwater, with enhancements made by [3] 4

1.3 Examples of the challenging perception environment underwater, with (a) attenuation (image courtesy of the McGill Mobile Robotics Lab) and (b) backscattering with additional occlusion due to bubbles and water quality. 5

1.4 Underwater gestures from an ocean trial. 6

2.1 Example skeletal joint mapping. 11

2.2 Demonstration of RoboChat Gestures from [4]. Image courtesy of the McGill Mobile Robotics Lab. 12

2.3 Examples of the token gestures implemented in [5]. 13

3.1 Sample hand positions for each of the gestures included in our study, (a) go somewhere, (b) pick up, (c) take a picture, and (d) general pointing. Images courtesy of Luoyao Chen. 17

3.2 Decision tree for baseline labeling policy. 23

3.3 Sample labeled images from our dataset. 24

3.4 A full example of the extended labeling policy. 25

3.5 EVA Labeling tool interface. Image courtesy of Luoyao Chen. 26

3.6 Training, validation, and testing dataset split. 27

3.7 SSD Architecture. CONV represents a convolutional layer while FC represents a fully-connected layer. Image courtesy of [6] 29

3.8 Faster R-CNN architecture. ConvNet represents the base feature extractor, while FCs are the fully connected layers. RoI stands for Region of Interest. Image courtesy of [6] 31
3.9 YOLOv3 Architecture, Image courtesy of [6] 32
4.1 Training loss of each of the networks (a) SSD, (b) Faster R-CNN as implemented in Detectron2, (c) YOLOv3, and (d) YOLOv5. Loss is plotted on the y-axis and number of epochs (YOLOv3, YOLOv5) or number of iterations (SSD, Faster R-CNN) is on the x-axis. 36
4.2 Demonstration of IOU scores, Image courtesy of [6] 38
4.3 Images with ground truth labels of “diver” which were misclassified as “diver_pointing.” 42
4.4 Images with ground truth labels of “diver_pointing” which were misclassified as “diver.” 43
A.1 Website part 1. 93
A.2 Website part 2. 94
A.3 Website part 3. 95
A.4 Website part 4. 96
A.5 Website part 5. 97
A.6 Intake form for research study 98

List of Abbreviations

ADRIATIC - Advancing Diver-Robot Interaction Capabilities

AI - Artificial Intelligence

AP - Average Precision

AUV - Autonomous Underwater Vehicle

BB - Bounding Box

CADDY - Cognitive Autonomous Diving Buddy

CNN - Convolutional Neural Network

COCO - Common Objects in Context

DTW - Dynamic Time Warping

FC - Fully Connected

FN - False Negative

FP - False Positive

FPS - Frames Per Second

GPU - Graphics Processing Unit

HMM - Hidden Markov Model

HRI - Human-Robot Interaction

ICP - Iterative Closest Point

IOU - Intersection Over Union

IRB - Institutional Review Board

IRV Lab - Interactive Robotic and Vision Lab

LoCO AUV - Low-Cost Open-Source Autonomous Underwater Vehicle

MLP - Multi-Layer Perceptron

R-CNN - Regional Convolutional Neural Network

RoI - Region of Interest

ROV - Remotely Operated Vehicle

RPN - Region Proposal Network

RRF - Random Regression Forest

SOP - Standard Operating Procedure

SOTA - State of the Art

SSD - Single-Shot Detector

SVM - Support Vector Machine

TP - True Positive

U-HRI - Underwater Human-Robot Interaction

VOC - Visual Object Classes

YOLO - You Only Look Once

Chapter 1

Introduction

In current scientific exploration, robots are the vanguard, taking us where no human has yet dared to venture. Prominent today is exploration in space, where the Perseverance rover and Ingenuity helicopter are exploring the surface of Mars [7]. However, there still remain new frontiers much closer to home. While much of Earth's *terra firma* has been traversed and mapped, the bodies of water covering over 70 percent of the surface of the earth remain largely unexplored [8]. Using robotics to explore aquatic environments can help us learn more about this domain to not only become better stewards of our planet, but also to work more safely where human activity intersects the aquatic domain. With the aquatic domain being by nature a challenging and potentially life-threatening environment for divers, developing autonomous underwater vehicles (AUVs) which can work collaboratively with divers in this space would greatly lower risk to human life, while extending our current capabilities. Developing an interaction framework where divers can instruct robots and transfer responsibilities for tasks that are difficult, dangerous, or undesirable for the divers to perform themselves is therefore a key component of research in underwater human-robot interaction (U-HRI) (Figure 1.1). In such a challenging environment for both diver and machine, we seek a fundamental communication system which is natural for both diver and robot; for divers, such a communication modality is already implemented in diver sign language [9]. Research in underwater human robot interaction has consequently taken inspiration from diver sign language to develop control systems for AUVs based on token gestures which are mapped to specific commands. To expand beyond these fundamental control systems, developing a gesture recognition system for more general and natural gestures seems a logical progression. Around the world, each culture appears to have some form of a pointing gesture to indicate direction, with its ubiquity making it considered a "building block of human communication" [10]. Thus we propose that developing a robust algorithm for recognizing pointing gestures underwater is a crucial step toward natural, effective, and intuitive human-robot interaction with autonomous underwater vehicles. This introduction serves to more fully ground and motivate this research topic.

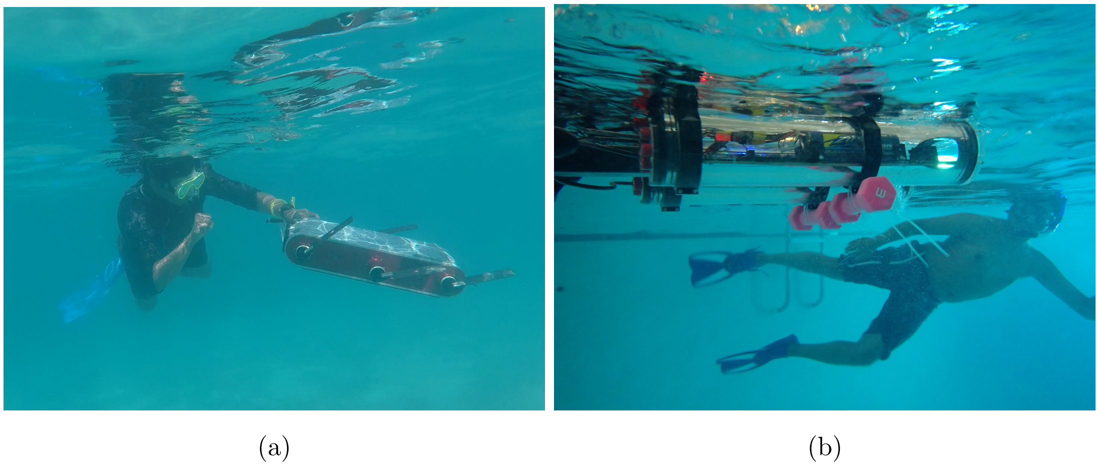

(a) (b)

Figure 1.1: Underwater human-robot interaction scenarios with the (a) AQUA robot [1] and (b) LoCO AUV [2].

1.1 Motivation for Underwater Robotics

Underwater robotics has many application areas supporting stewardship of our planet, conservation efforts, industrial activities, and scientific exploration. By the nature of the water cycle, trash is being swept into our rivers, and ultimately our lakes and rivers. This problem scales quickly when non-biodegradable material contaminates our waterways and affects ecosystems [11]. Underwater robotics has already demonstrated the capability of visually detecting and identifying underwater trash, a first step towards the ultimate goal of "exploration, mapping, and extraction of such debris by using AUVs" [12]. Beyond underwater trash, our oceans also contain WWII munitions and unexploded ordinance, which can be identified and potentially disarmed by underwater vehicles [13] [14]. In the area of conservation, there are plentiful opportunities for underwater robotics to support wildlife monitoring, as well as track the spread of invasive species [15].

Underwater robotics can also be applied in areas with more natural room for human-robot interaction such as industrial applications, search and rescue missions, and scientific exploration. Mueller *et al.* have already demonstrated diver-robot collaborative ventures for underwater bridge inspection for damage post-flood disasters [16], [17]. Further, ship inspection, a dangerous activity primarily performed by divers, is increasingly being performed with the aid of AUVs [18]. Protocols for underwater search and rescue, an inherently dangerous yet crucial task, are currently under development [17]. Even underwater archaeological missions [19] and marine biology surveys [20] are using underwater vehicles to support their work.

Inherently an adverse environment for both man and machine, the aquatic domain has constraints which present unique challenges. Divers are constrained by the physical limitations of the human body to handle the pressure of water at depth, thermoregulation in the aquatic environment, and dependence on an external oxygen supply for respiration underwater. This makes diving an inherently dangerous activity. For an underwater vehicle, the attenuation of electromagnetic signals and diffusion of sound waves render these traditional communication modalities highly ineffective. Machine vision-based communication methods are thus a natural choice for underwater perception for AUVs; however, this modality is also challenged by the unusual optical effects of light in water, namely refraction, backscattering, and attenuation [21], as well as limited visibility due to water quality (Figures 1.2 and 1.3). Thus the challenge of creating autonomous underwater vehicles (AUVs) capable of assisting divers in performing difficult, dangerous, or undesirable tasks is a central motivation for this work.

1.2 Human-Robot Interaction for Autonomous Vehicles

In order for a diver and an AUV to work together, there must be an established interaction protocol. Study of such protocols is the focus of the field of human-robot interaction (HRI). Our specific area of interest is human-robot interaction with autonomous vehicles in field robotics; this domain is particularly challenging due to the inherent variability

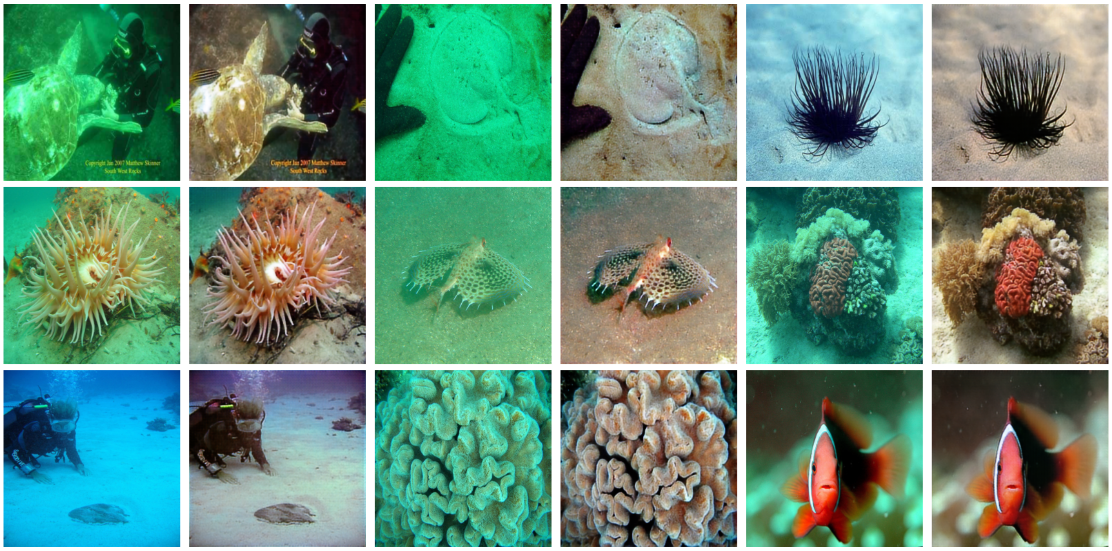
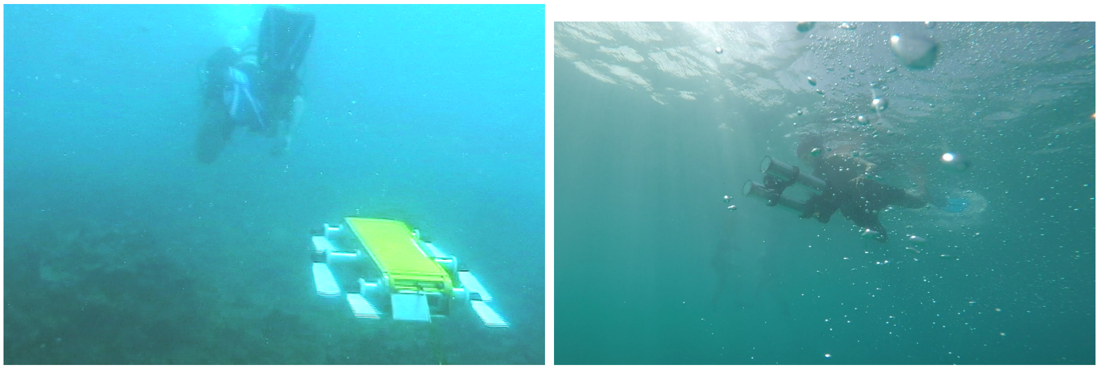

Copyright Jan 2007 Matthew Skinner
South West Rocks
Copyright Jan 2007 Matthew Skinner
South West Rocks

Figure 1.2: Examples of the challenging, visually degraded imagery faced underwater, with enhancements made by [3]

of environmental conditions and the wide range of potentially desirable interaction capabilities. Due to these obstacles, human-robot communication modalities for working with autonomous vehicles is a widely explored area.

In the terrestrial realm, communication between humans and autonomous vehicles can take multiple forms, depending on the application. With the voice being the most natural communication vector from person-to-person, autonomous vehicles have been equipped with microphones and auditory processing algorithms, enabling instruction through vocal commands [22]. Beyond sound, person-following robots have also utilized haptic technology for human control, with signals received through force sensors [23] [24]. Moving away from direct auditory or tactile interactions, a remote control smartphone interface has even been introduced for controlling autonomous luggage [25]. Beyond these, gestures are yet another communication modality for robustly interacting with an autonomous vehicle, with the signals being captured by either sonar or traditional monocular or stereo cameras [26] [27].

Given these varied options for human-robot interaction, due to the extreme attenuation of electromagnetic signals and distortion of sound underwater, the most natural modality for underwater human-robot interaction is through machine vision with a



(a)

(b)

Figure 1.3: Examples of the challenging perception environment underwater, with (a) attenuation (image courtesy of the McGill Mobile Robotics Lab) and (b) backscattering with additional occlusion due to bubbles and water quality.

monocular or stereo camera.

1.3 Gestural-Based Communication

When using monocular or stereo cameras as a primary means of perception, gestures become a natural means of control for autonomous vehicles. In their survey of gesture recognition algorithms, Mitra *et al.* identify three broad categories of gestures: hand and arm gestures, head and face gestures, and full-body gestures [28]. Each of these classes of gestures has been used to direct autonomous vehicles. Hand and arm gestures naturally comprise many instructions, not the least among them being pointing gestures. Head and face gestures can convey agreement or dissent through a head shake or nod [29], while gaze has even been used to single out individual robots for further instruction [30]. Full-body gestures allow the relative position of the arms to the rest of the body to become the basis for gestures [31]. All of these categories of gestures together can form a robust communication system, as demonstrated by Canal *et al.* in [29], where they develop a gesture recognition system for humanoid robotic assistants.

In the aquatic domain, gestures are currently the central means of communication between divers and autonomous vehicles for various reasons (Figure 1.4). The aforementioned challenges of the underwater domain lead machine vision to be the primary mode of perception; however, this is particularly advantageous for additional reasons. Divers controlling an autonomous vehicle underwater are already under heavy cognitive load without the introduction of a robot, balancing tasks such as monitoring their dive gear, maintaining proper depth, and managing their orientation in a six degrees-of-freedom environment. These critical responsibilities require a baseline level of a diver's attention, which is further occupied with accomplishing the dive mission. As autonomous vehicles are intended to aid divers rather than put additional strain on their cognitive abilities, creating the most natural communication interface between autonomous underwater vehicle (AUV) and diver is critical. With gestures being a fundamentally familiar non-verbal means of communication, they are a logical choice for control of AUVs.

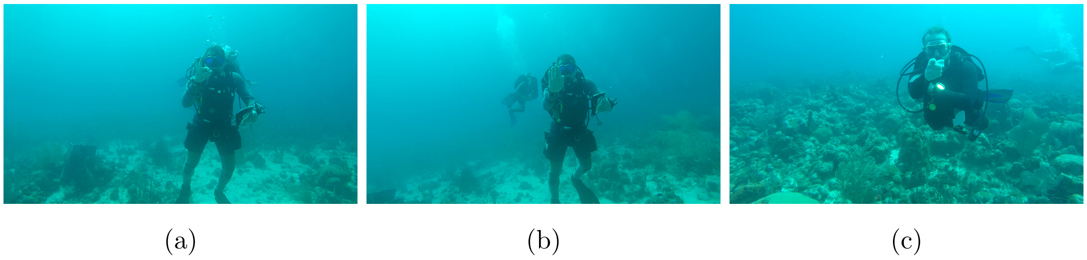


Figure 1.4: Underwater gestures from an ocean trial.

1.4 Pointing Gestures

Pointing gestures are a key nonverbal communication modality in collaborative scenarios, fundamentally conveying the directionality associated with an interaction or command. Consider the phrases "the cafeteria is down that hallway," "please bring me that box," "exit through that door," and "look at that bird"; in a real-life scenario, each of these short phrases would naturally be accompanied by a pointing gesture to direct the attention of the agent toward the correct hallway, box, door, or bird. These sample interactions likewise demonstrate some of the intents associated with pointing gestures: to explore an area, retrieve an object, follow a specific path, or observe (in the case of an autonomous vehicle, take a picture of) an item. With directionality derived from pointing gestures being crucial in collaborative tasks, recognizing pointing gestures is a fundamental capability of AUVs. Further, robust gesture recognition and directional inference is key for autonomous vehicles, since there is no explicit human-in-the-loop control mechanism. Unlike remotely operated vehicles (ROVs), AUVs do not inherently involve a human-in-the-loop to correct misinterpreted commands, which makes the ability of AUVs to independently and accurately interpret commands both mission- and safety-critical. This is of even greater importance underwater, where risks due to error are elevated simply due to the nature of the aquatic domain. Thus we conclude that the development of a robust algorithm for pointing gesture recognition is a compelling research topic for advancing the collaborative capabilities of autonomous vehicles in underwater human-robot interaction scenarios.

1.5 Contributions

This thesis provides the following contributions:

- A study focused on the modality of humans pointing to objects.
- Two fully annotated pointing datasets, one fully in the underwater domain which contains the classes of diver and diver_pointing, and a second in the terrestrial domain.
- An analysis of the efficacy of four state-of-the-art object detection networks in identifying pointing gestures in the underwater domain.

The remainder of this thesis is organized as follows: Chapter 2 contains a review of the relevant work in the areas of human-robot interaction, general and pointing gestural communication in robotics, and gestural communication in underwater robotics. Chapter 3 outlines the methodologies followed, including the human study design, annotation and generation of the datasets, and modification of the object detection networks for work with custom datasets. In Chapter 4, the experiments training the object detectors with these datasets and the results of these experiments are discussed. Finally, conclusions and thoughts on future directions from this work are discussed in Chapter 5. Supplementary material follows in the appendix.

Chapter 2

Related Works

2.1 Human-Robot Interaction

The area of Human-Robot Interaction (HRI) is a broad field, with different aspects of research focusing on perception [29], motion planning [32], understanding human-robot trust [33], and developing human-robot collaborative protocols [34]. Since our work specifically focuses on robotic perception of gestures as a human-robot communication modality, we review the related work in the areas of terrestrial and underwater robotic gesture recognition in more detail.

2.2 Gestural-Based HRI

Gestures in Robotics

Both full-body gestures as well as hand gestures have been robustly investigated in robotic control, with various algorithms underpinning the gestural recognition. Early work in this area utilized parametric Hidden Markov models (HMMs), which had the advantage of being temporally invariant [35]. Extending this idea, Nickel *et al.* track a person's face and hands, and use a trained HMM to recognize when a pointing gesture occurs [36]. After recognizing the gesture, the authors further infer the pointing direction in three ways: utilizing a line from head to hand, extending the forearm, and estimating head orientation. The whole-body nature of gestural communication is evidenced by the work of Couture *et al.*, who do not restrict gestures to the arms and hands, but rather introduce a methodology for commanding individual robots from a multi-robot system by first selecting a robot through a gaze and subsequently assigning it a task with an arm gesture [30]. Specifically focusing on humanoid robots for human assistants, the authors of [29] recognize two classes of gestures, static and dynamic, creating a multi-robot system capable of recognizing a hand wave, pointing gesture, head shake, and nod. By extracting the skeletal joint features of a person, they split their approach to use Dynamic Time Warping (DTW) [31] for dynamic gesture recognition, while analyzing the geometrical pose of the joints over a time series of frames to identify a static point. These implementations demonstrate a portion of the wide array of gestural systems for robotic control.

With the extensive work investigating gestures for robotic control, numerous datasets have been released to support advancement of such gestural recognition systems. However, among these datasets, a focus on pointing gestures is notably absent. Creating a multimedia dataset for human-robot interaction including both audio and visual data from a Kinect sensor system, the authors of [37] address a pointing-relevant intent, *i.e.*, a gesture to "go somewhere," yet don't explicitly assign a gesture for pointing. Likewise, the ChaLearn Looking at People project [38] has created multiple datasets for human pose estimation and gesture recognition, but none specifically analyzing pointing gestures. Similarly, a pointing gesture is notably missing from both the 29-class Praxis upper body dataset [39] and the DVS128 Gesture dataset of 11 hand gestures released by IBM Research [40]. While conducting a broader survey of datasets for human gesture recognition which focused on hand and arm movement, the authors of [41] observe that in general, datasets primarily focus on vocabularies of learned gestures such as sign language [42], military gestures [43], or cultural signs [44]. In contrast, the datasets in this work have been created specifically to support recognition of the natural, universally understood pointing gesture.

Pointing Gestures in Robotics

The task of identifying pointing gestures has been widely explored in terrestrial robotics, with algorithms varying from a focus on hand pose to a full body approach, utilizing either stereo or monocular camera input. Beginning with hand pose algorithms, Fujita *et al.* focus specifically on identifying the pose of a hand pointing toward a forward-facing camera [45]. To do so, they utilize two cameras with parallel optical axes placed several meters from the hand and process the pair of images using a Random Regression Forest with stereo techniques. They further extend accuracy by utilizing Bayesian techniques over sequential frames. Also focusing on the hand position, the authors of [46] instead employ a probabilistic approach to estimating pointing gestures, which is notably independently of body pose. Shifting focus to algorithms leveraging full-body pose, [47] utilizes a Kinect sensor to identify the 3D skeletal joint mapping of the human instructing the robot. Making the simplifying assumption that the raised arm is performing a pointing gesture, they use only the hand and shoulder coordinates to extract the pointing vector direction, subsequently performing a 3D to 2D mapping so that the robot can interpret the 3D pointing gesture as a cardinal direction. Also inferring the pointing gesture from a 3D skeleton (Figure 2.1) obtained from a Kinect sensor, the authors in [48] further present a method of calibrating their pointing gesture recognition algorithm with horizontal and vertical offsets to compensate for users' natural variation in pointing gestures. Taking these algorithms a step further, [49] and [50] extend the pointing direction as a vector emanating from the elbow coordinates through the wrist coordinates. They subsequently identify the robotic navigation destination as the intersection of this vector with the floor plane. Each of these approaches inherently uses depth information in conjunction with stereo cameras.

Additional work in identifying pointing gestures using monocular cameras has also found success, frequently paired with a machine learning algorithm. In [51] and [52] the authors use semantic segmentation to identify a human, and the gesture recognition problem is solved by training a support vector machine (SVM) classifier. Also using a monocular camera, [53] tracks both the hands and the face, recognizing a pointing gesture based on finger pose and subsequently estimating the object being pointed toward through both the face and hand orientation. In the context of a "robot service companion," Richarz *et al.* develop an algorithm for pointing gesture recognition using monocular cameras, limiting the interaction space to 2 meters between human and robot [54]. They define the pointing gesture rigidly as an extended arm, with the head and gaze aligned in the intended pointing direction, and utilize a hierarchical neural classifier based on multi-layer perceptrons (MLPs) to estimate the radius and angle defining the target location. This work is extended upon in [55]. With such a large body of work dedicated to pointing gestures in the terrestrial domain, the groundwork has been laid for developing such algorithms in the underwater domain.

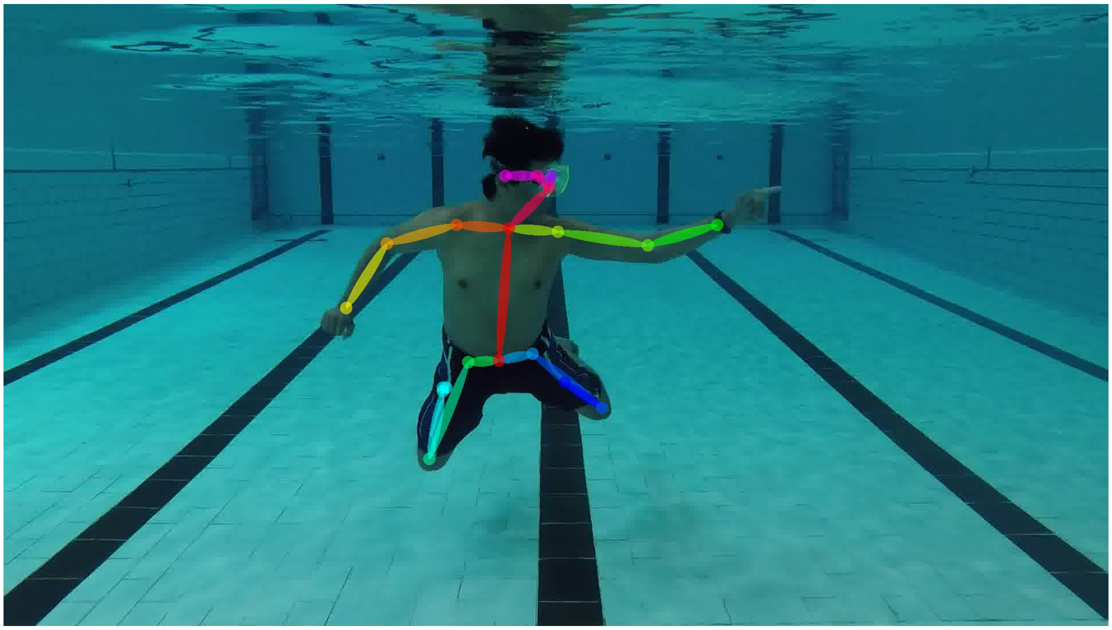


Figure 2.1: Example skeletal joint mapping.

2.3 Gestural Communication in Underwater Robotics

While the pointing gesture recognition and inference problem has been widely investigated in terrestrial robotics using machine vision, it has yet to be addressed in the underwater domain. Thus the related work in the underwater domain primarily focuses on the more general problem of identifying gestures made by divers. Early work in this domain by Dudek *et al.* utilized fiducial markers, with a gesture consisting of a diver displaying a single marker [56]. With a meaning mapped to each fiducial marker created using the ARTag toolkit [57], sequences of gestures made with the tag compose a language, called RoboChat. The authors extend this schema in [4] to create RoboChat gestures, a language involving only two fiducial markers, where the relative motion, *i.e.*, physical gesture performed with the two markers, defines a command (Figure 2.2). With one marker serving as a reference point, the point cloud extracted from the motion of the second marker is compared to the shape of a known gesture using an Iterative Closest Point (ICP) algorithm [58]. This early work in the underwater gesture recognition laid the foundation for future development.

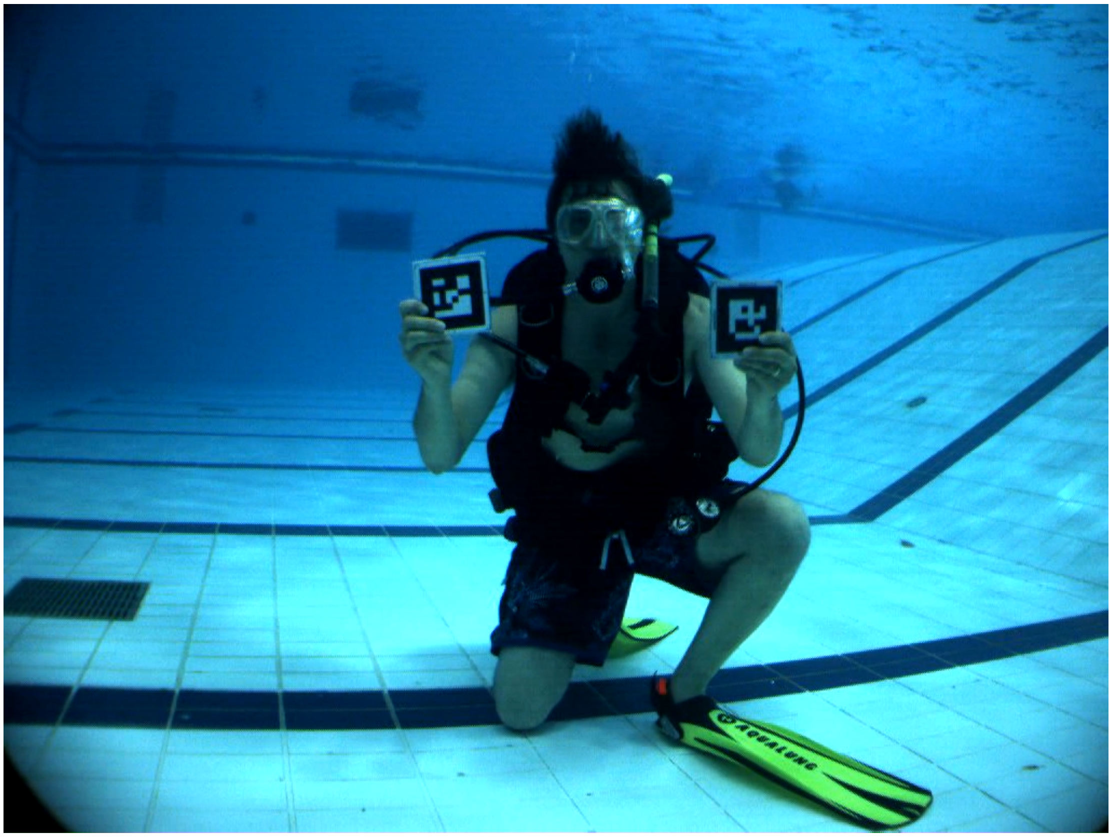
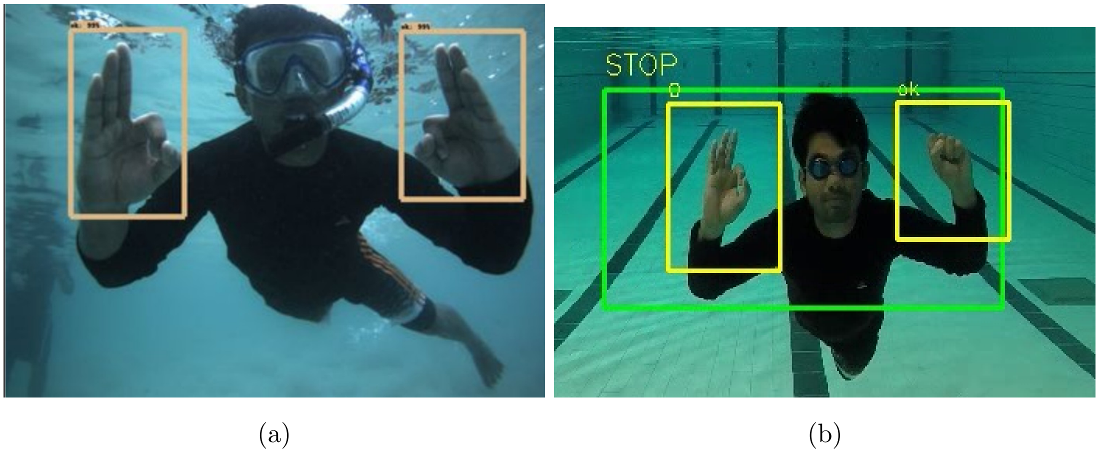


Figure 2.2: Demonstration of RoboChat Gestures from [4]. Image courtesy of the McGill Mobile Robotics Lab.

More recent work in underwater gesture recognition focuses on recognizing gestures without supplementary vision aids like markers. The Cognitive Autonomous Diving Buddy (CADDY) project investigated the development of a gesture-based language for controlling an AUV [59], which was initially implemented using a Haar Cascade classifier to identify gesture candidates, with validation and classification performed by MultiDescriptor Random Forests [60]. A broader gestural recognition system implementation is demonstrated in [61], based on the CADDIAN gestural language, which is described in detail in [62] and [17]; however, in these works the gesture recognition system backend is not discussed. Beyond the CADDY project, numerous other groups are investigating gesture recognition. Islam *et al.* designed a gestural control interface for the AQUA robot [1] where specific gestures are mapped to token commands, with sequences of these commands being interpretable by the robot as instructions. In this work, the gestural recognition system is implemented using a deep approach combining a region proposal network (RPN), based on contour qualities such as the convex hull, with a convolutional neural network (CNN) to perform the gesture classification task [5]. Also using deep learning, Mital *et al.* seek to classify diver hand signals [9], beginning with images of a hand only (eliminating the need for a region proposal network). They focus on extracting eight Hu moment parameters from a background-subtracted image, ultimately feeding these eight parameters into an artificial neural network for classification [63]. Diverging from the deep approaches, the authors of “Development of an Underwater Hand Gesture Recognition System” take a more fundamental computer vision approach, preprocessing the image to obtain a segmentation of the diver’s hand, defining a wrist line by fitting circles to the hand and palm regions, and extracting the actual hand pose after rotating and cropping the image. The gesture, *i.e.*, the specific hand pose, was extracted using two methods, one using a convex hull, and the second with finger segmentation [64].



(a)

(b)

Figure 2.3: Examples of the token gestures implemented in [5].

Other work in the area of underwater gesture recognition diverge from the use of traditional vision methods for gesture recognition. Building off the work started by the CADDY project, the “Advancing Diver-Robot Interaction Capabilities” (ADRIATIC) project utilizes a diving glove with motion and tensile sensors to capture inertial movement which is used to detect and classify gestures [65]. More closely related to the vision approaches previously discussed, the CADDY project also investigated using high resolution multibeam sonars, also referred to as acoustic cameras, to perform hand gesture recognition using three methods: a convex hull approach, a support vector machine method, and finally a novel algorithm combining both a convex hull and a SVM [66].

Among the underwater gesture recognition protocols discussed here, the majority require a close proximity of the diver to the AUV, with the gestural languages mostly concerned with the hand pose. Further, none focus specifically on pointing gestures. These two observations reveal the unique standing of this work among the current literature: recognition of pointing gestures through utilizing a full-body diver pose.

Chapter 3

Methodology

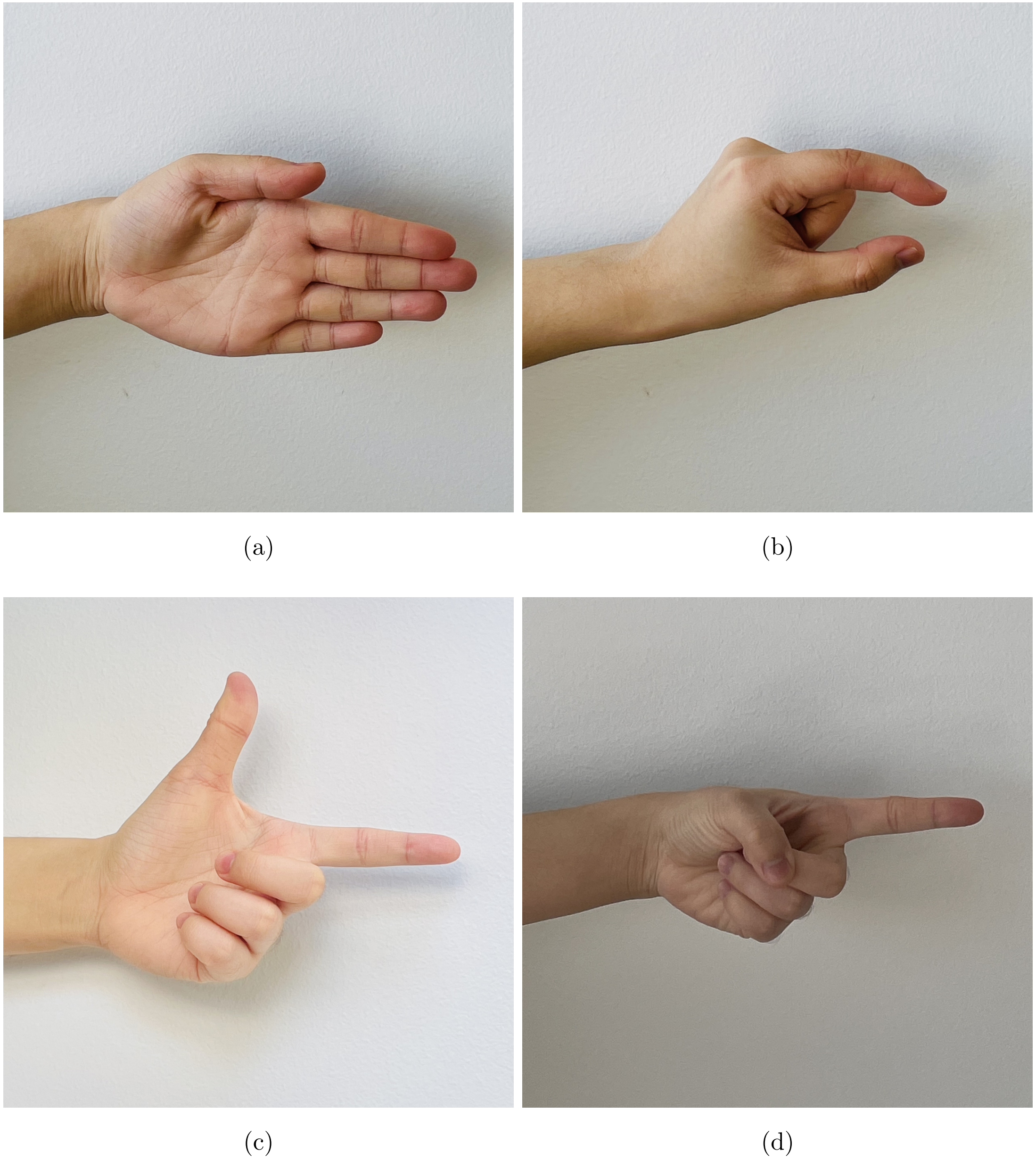
Since we have chosen to use a deep learning approach to developing a pointing gestural recognition system for underwater HRI, a key aspect of this work was developing a large dataset necessary for training deep neural networks. Due to the majority of this work being completed during the ongoing novel Coronavirus (COVID-19) pandemic, data collection opportunities were limited due to suspension of our research laboratory's in-person collaborative activities. Specifically, prior to the pandemic we performed pool trials one to two times a month during the academic year, collecting data and testing our algorithms during development. In the summer months, this frequency increased, with our lab taking advantage of the warmer temperatures to collect data in the field, testing our algorithms in Minnesota lakes. A yearly trip to Barbados afforded further validation during ocean trials. These data collection and verification opportunities were all placed on hold during the pandemic. Thus we relied upon collected data from previous trials and also designed a study to collect data of participants performing pointing gestures, which was submitted to the University of Minnesota's Institutional Review Board (IRB) for approval due to the involvement of human participants. With the study approved and data obtained, we both developed and executed a labeling scheme to annotate this data for use in training object detection networks to solve the pointing gesture problem. Once the dataset was prepared, it was used to train four different state-of-the-art object detection networks for comparison of their efficacy in pointing gesture recognition in the underwater domain. The sections below expand on each of these facets of this research.

3.1 Human Study

Study Design

Thousands of images are necessary to train deep models, and since no publicly available dataset specific to pointing gestures exists, we designed a human study to develop such a dataset and understand better how people point. Since our goal is to create a natural interaction between human and robot, we designed tasks that would encourage natural movements, while constraining the form of the pointing gesture to specific hand positions to define reproducible (and thus machine learnable) representations. Defining the hand positions classified as "pointing" also permits the correlation of intent to the modality of pointing, an area of investigation beyond the scope of this thesis' contribution, but currently under investigation by our collaborator Luoyao Chen. Lastly, the recording setting is specified to help support domain-independent representation learning in the deep networks.

With the pointing gesture being a ubiquitous motion with variation around the world, we deliberately decided to specify the hand positions for our participants to assume when making a pointing gesture. The motivation for this specification was two-fold: to standardize what is considered a pointing gesture for our algorithm, and to facilitate a more robust communication methodology between human and robot through inferring diver intent from the form of the pointing gesture used. Accordingly, four different classes of pointing gestures were defined, shown in Figure 3.1. Each of these four classes are a variation on the traditional pointing gesture. The first, with all fingers extended and held together with the palm facing outward was designated for the intent of "go somewhere" (Figure 3.1 (a)). The second, with the index finger and thumb extended in parallel was designed to indicate a "pick up" command (Figure 3.1 (b)). The third and fourth pointing gestures both have the index finger extended; however the former extends the thumb upwards at a right angle to the index finger to indicate "take a picture" (Figure 3.1 (c)), while the latter folds the thumb over the remaining curled digits, forming the classic index finger point (Figure 3.1 (d)). In the study we provided reference images of these classes of gestures, with the clarification that they may be made in any orientation, and also featured two demonstration videos. This provided a baseline expectation for the valid pointing gestures in our research study.



(a)

(b)

(c)

(d)

Figure 3.1: Sample hand positions for each of the gestures included in our study, (a) go somewhere, (b) pick up, (c) take a picture, and (d) general pointing. Images courtesy of Luoyao Chen.

In addition to defining the hand positions for the pointing gestures, we designed the research study around specific sets of activities. By giving participants an activity during which to perform the gestures, we hoped to encourage natural pointing gestures by incorporating them into familiar tasks and by diverting participants' focus from the camera to the actions being performed. Sample activities for the "go somewhere" gesture included directing people to go to a location during a tour, and telling a pet to go outside. Similarly, we suggested that the "pick up" gesture be performed in a scenario where the volunteer is cooking or making a craft with an assistant who brings the necessary utensils. Since it is more easily incorporated into general activities, we simply suggested that our participants use the "take a picture" gesture when pointing to an object. Lastly, to encourage exploration in different pointing scenarios, we defined pantomiming a weather report and "any other activity that involves making natural pointing gestures" as valid activities for the video submissions. These sets of activities were designed to guide participants to make natural pointing gestures in the specified hand positions.

Lastly, we placed some constraints the participants' video settings. To ensure that most of the participants' frame is in the field of view of the camera, we requested that our participants place themselves a moderate distance from the camera, approximately two meters. In addition, we suggested uncluttered backgrounds, to help prevent our deep networks from learning features from the scene surrounding the person pointing. Further, we requested that the videos not utilize any filters or color-correction, aiming for natural lighting and undistorted representations. Finally, we requested that the videos not contain geo-location data or audio, to help protect our participants' privacy. These video settings were imposed to support the research purpose of studying natural gestures as would be captured by a robotic perception system.

IRB Human Studies Regulatory Review Process

Since our research study involves humans participants, it is subject to review by the Institutional Review Board (IRB) of the University of Minnesota, which is an advisory board overseeing compliance with Federal regulation and University ethical policy. Federal regulation of human studies is grounded in principles identified by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, which was created in 1974 to identify fundamental ethical principles to guide human studies. The Commission published its findings in the Belmont Report [67] in 1976, presenting three overarching principles: Respect for Persons, Beneficence, and Justice. Federal regulation centers around these three ethical pillars. In addition to ensuring compliance with Federal regulations, the IRB also ensures that human studies follow the guidelines set in place by the University. Designing our study around these principles, we accordingly submitted our research study proposal to the IRB for review.

Making an initial research study proposal to the IRB required extensive planning and documentation of our intended study operational procedures. Central to this was completing a standard document, HRP-580 Social Template Protocol (Appendix A.1). In this protocol the principal and student investigators are identified, along with the objective, significance, end goals, and procedures of the study, as described above. To help protect the privacy of the participants, data management practices, specifically data anonymization, storage handling, and access restrictions are also outlined in the protocol. Further documentation includes the study duration, included populations (to ensure protection of vulnerable groups), participant withdrawal procedures, and potential risks to participants, all designed to ensure participants' physical, mental, and emotional safety. Likewise, study confidentiality and compensation proposals are prepared and reviewed within this document. Central to the IRB submission, this protocol dictates the study standard operating procedures (SOP).

In addition to the study protocol, the initial IRB proposal requires submission of any supplementary materials relevant to the study. For our study, this included a volunteer consent form and the recruitment promotional materials. Similar to the Social Template Protocol, the consent form is a standard document, HRP-582 Social Behavioral Consent Form (Appendix A.2). Within this document, an outline of the study goal and the participants' role in the study is presented in a non-technical manner. Specifically, the motivation of the research study, its duration and scope, the risks involved with participation, volunteer responsibilities, compensation details, post-study data banking, and withdrawal processes are presented prior to requesting participants' written consent to participate in the study. Beyond this consent form, the data collection form (Appendix A.7), as well as the promotional materials for the study (including the content for a web landing page, social media promotional posts, and a distribution flyer) are all submitted as supplementary materials (Appendix A.6).

A final component of the initial IRB study proposal involved each of the investigators listed completing training relevant to ethics and conducting research involving human participants. The two required courses, Research Involving Human Subjects (RCR) and Social / Behavior or Humanist Research Investigators and Key Personnel, familiarize the researcher with the Belmont Report [67] and its core principles, as well as additional ethical considerations and approaches for conducting research involving humans. We take these principles seriously, and in adherance to the protocols set in place for our research study, the data collected remains internal to the IRV Lab. Thus the sample images here are not from our study, but have been created for demonstration purposes.

After the initial IRB review, our study was determined to be exempt from IRB review (Appendix A.3). This means that while we continue to adhere to the operating procedures outlined in the Social Template Protocol, the Consent Form is no longer necessary and has been replaced with an Information Form (Appendix A.4). Once we obtained final approval (Appendix A.5) from the IRB, we began our study, which lasted from February through May 2021. Summary statistics from this study may be seen in Table 3.1.

Statistic	**Value**
Number of Participants	31
Number of Video Submissions	39
Seconds of Video	1960
Minutes of Video	32.6
Approximate Number of Frames	58,800

Table 3.1: General Statistics on the Study Data Collected.

3.2 Dataset Preparation

Data Sources

The datasets utilized for training and evaluating the models in this work consisted of series of frames extracted from video clips. There were three main sources for these videos: past Interactive Robotic and Vision Lab (IRV Lab) pool and field trials, public domain footage from the National Park Service's B-Roll archive [68], and volunteer submissions to the research study described in §3.1. The primary data used for training the networks consisted of pool trial data: video of divers collected underwater using monocular GoPRO cameras [69] and the AQUA robot's stereo cameras [1]. The last two data sources were video taken in terrestrial settings.

Labeling Policy

With initial data collected, the next step is to annotate the data. In deep learning, the nature of the algorithm being developed informs the annotation approach. Since we are interested in the twofold problem of localizing a diver within a frame and also identifying the gesture being performed by that diver, our algorithms fall into the class of networks known as object detectors. Training an end-to-end object detection framework requires supervised learning, and therefore our annotated dataset consists of frames with object instances labeled using bounding boxes. Prior to beginning this labeling task we defined annotation rules, which we built into an overall labeling policy to guide and standardize the annotation process. The labeling policy for the pool trial data (described in section 3.2) served as a baseline policy, and this policy was extended for the study data.

The baseline labeling policy for these datasets addressed two main questions: what should be labeled in each frame, and how do we differentiate between the two classes, pointing and non-pointing. Our first set of annotation rules governs the first question, "what should be labeled in the video frames?" For the baseline policy we decided to restrict our annotations to drawing bounding boxes around three classes of objects: a diver, a diver_pointing, and (if applicable) the object_indicated (by the diver_pointing). Any other objects in the frame are extraneous to the gesture recognition we seek to learn, so will not be annotated in the dataset. Within these classes, each instance of these three classes is labeled unless the class instance becomes fully occluded in a frame (for example, a diver passing behind a coral reef which completely obscures them from view). Secondarily, we draw a bounding box around the entire diver (whether in the diver or diver_pointing class), additionally encompassing their dive gear if applicable. If the diver is partially occluded or only partly within the video frame, the diver is labeled if the majority of their body is visible; in this case the bounding box should encompass the entire diver, including occluded limbs or as much of the diver as is within the frame.

These annotation rules address the first question of what we label in our dataset.

A second set of annotation rules define our process for how to differentiate between the two classes of diver, a diver and a diver_pointing. Since the action of making a pointing gesture is a fluid motion, it can be challenging and subjective to determine at which frames a pointing gesture begins and ends. We decided to define the labeling of a pointing gesture based on two items: hand position and arm rigidity. To be classified as a frame with a diver_pointing, the diver's hand must be in one of two positions: the traditional point, with index finger extended (Figure 3.1(d)), or the full-hand indication, with all five fingers together and pointing in a single direction (Figure 3.1(a)). In addition, the diver's arm must be either fully extended or held rigidly in the pointing gesture (in an either extended or non-extended state). To clarify uncertain cases, a labeling decision tree was created, as shown in Figure 3.2. These two annotation rules together define whether a diver detected in a video frame is classified as a diver or diver_pointing. These are the rules of our baseline labeling policy. Sample labeled images may be found in Figure 3.3.

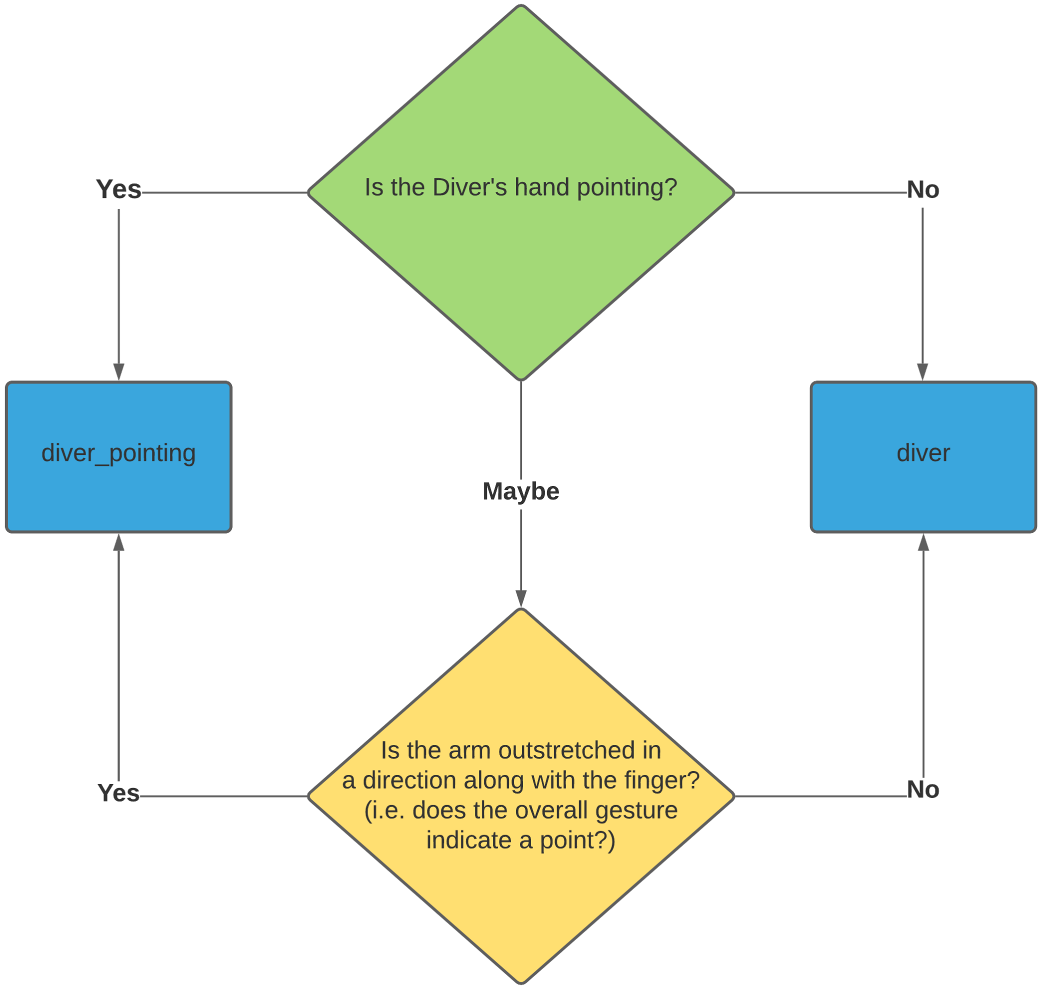
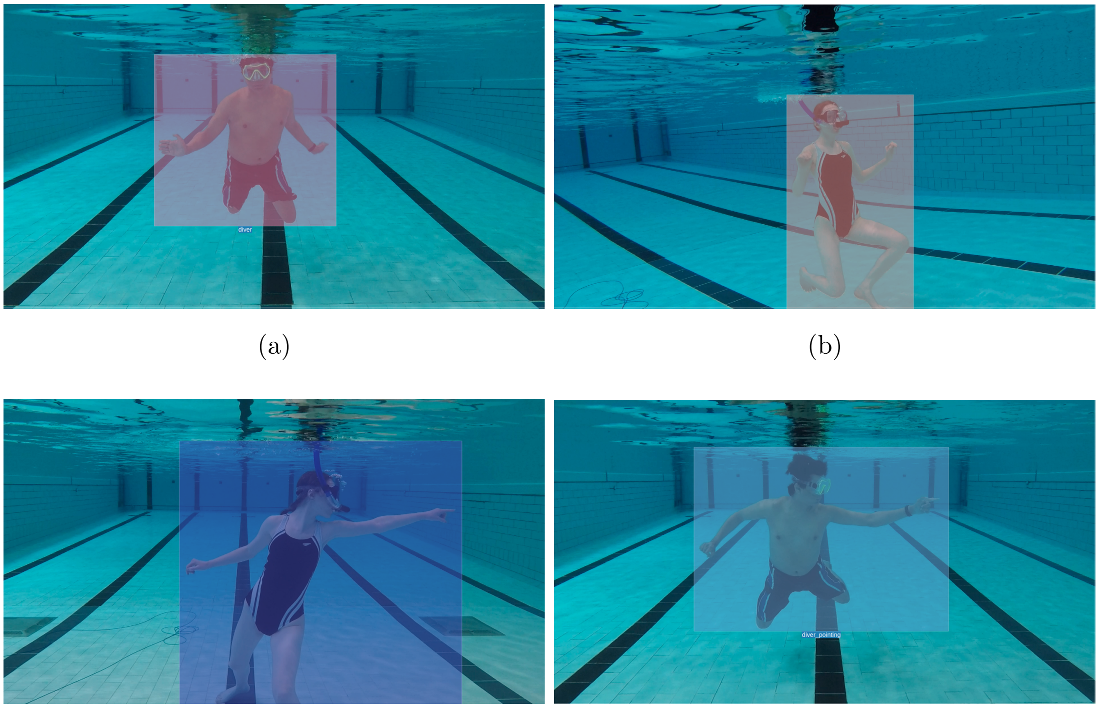

Is the Diver's hand pointing?
Yes
No
diver_pointing
diver
Maybe
Is the arm outstretched in a direction along with the finger? (i.e. does the overall gesture indicate a point?)
Yes
No

Figure 3.2: Decision tree for baseline labeling policy.

The baseline labeling policy was applied to both the pool trial data and the public domain videos; however, an extended labeling policy was used for the data from the study described in §3.1. For this dataset, we retain the diver, diver_pointing, and object_indicated labels, despite the study data being terrestrial. The rules to determine whether to label a person as diver or diver_pointing can be summarized as follows:

- A person who is not pointing is labeled diver



(a)

(b)

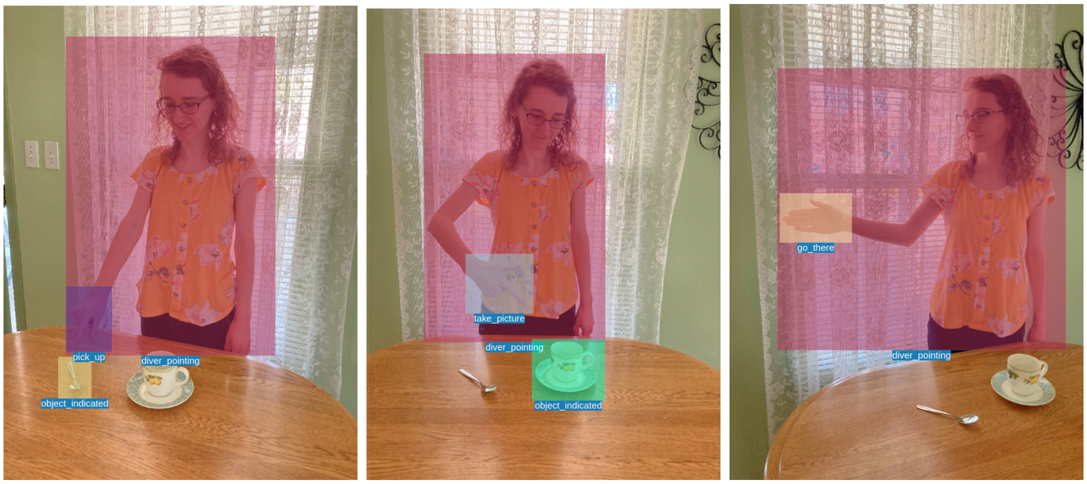
(c)

(d)

Figure 3.3: Sample labeled images from our dataset.

- A person who is pointing will be labeled diver_pointing
- In addition to the two pointing gestures defined in the baseline labeling policy, consider all four hand positions in Figure 3.1 as a pointing gestures, requiring the overall bounding box for the person to be diver_pointing
- In cases of uncertain hand position, follow the decision tree (Figure 3.2 from the baseline labeling policy) to determine whether to label as diver or diver_pointing

This labeling policy extends the baseline labeling policy by requiring up to two additional labels for any frame which is labeled as diver_pointing. Whenever a person is labeled as diver_pointing in a frame, also label the hand that is pointing, and if applicable, the object being indicated by the pointing gesture. When drawing a box around the hand that is pointing, this box should have one of four additional labels, based on the hand position assumed from 3.1 : (a) go_there, (b) pick_up, (c) take_picture, and (d) general_point. Lastly, if the pointing gesture is directed toward an object that is visible in the frame, also draw a bounding box around the entirety of the object with the label object_indicated. A full example of this extended labeling policy is in Figure 3.4.



(a) (b) (c)

Figure 3.4: A full example of the extended labeling policy.

Labeling Tool

To complete the labeling task, we used an open-source annotation tool hosted by the IRV Lab called EVA [70]. EVA is specifically designed to aid in creating annotations for object detection, providing a framework for drawing bounding boxes with class labels on the frames extracted from uploaded video clips or image sequences. To set up a labeling job within EVA, we first define the class labels, then create a project, to which the relevant class labels are added, assigning a number to each class in the process. Next we upload our videos or frame sequences to the project in EVA, where they are separated into 100-frame segments. We then draw a bounding box around each object in the first frame, and invoke the tracking feature in EVA. This tracking feature automatically predicts the bounding box for each object in the subsequent frames, which can then be adjusted for accuracy. This greatly speeds up the labeling process, and when an entire video has been labeled, the annotations can be exported from EVA in YOLO or PASCAL VOC format via a simple download. The EVA Labeling interface is shown in Figure 3.5.

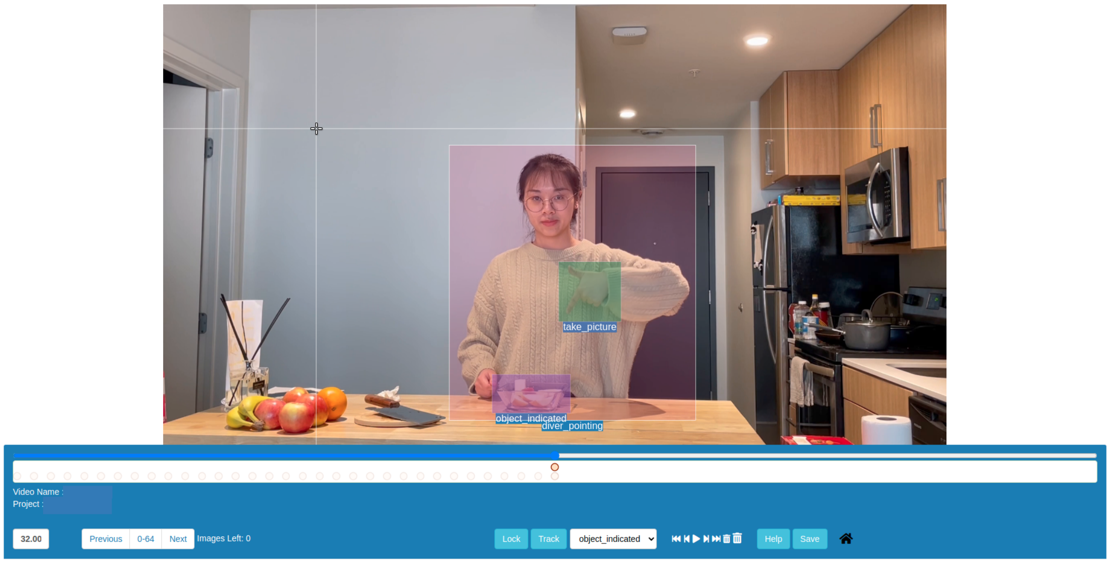


Figure 3.5: EVA Labeling tool interface. Image courtesy of Luoyao Chen.

Dataset Generation

After the labeling was completed with EVA, the annotations were reviewed a second time as a proofing step prior to exporting both the frames and annotations from EVA. This proofing step is crucial, because during this step we ensure both that all the class labels adhere to the labeling policy and that every frame is labeled, two points critical to success in supervised learning. After the data export, the dataset is cleaned, with two classes of frames removed: out of domain frames and blurry frames. Specifically, out of domain frames refer to pool trial data where the camera is above water, or frames where there is no diver present. Blurry frames are selectively removed from the dataset, dependent on the level of blur and the class label of the annotations in the frame. For frames labeled diver_pointing, the frame is removed if the level of blur is high enough that the pointing gesture, specifically the hand, is indiscernible. For all other frames, the frame is removed only if the level of blur makes it unfeasible to discern the general features of a diver. Once the dataset has been cleaned in this manner, the data is split into training, validation, and testing sets.

For this work we created several versions of our pool trial dataset, each following the standard 70% - 10% - 20% training, validation, and test split. Originally we split each video in the dataset temporally, with the first 70% of the frames for training, the next 10% for validation, and the last 20% for test. However, this yielded a highly imbalanced dataset where the training and validation sets had diver_pointing class instance rates in the 30th percentile, while the training set only contained a proportion of diver_pointing in the tenth percentile. To resolve this, two additional dataset splits were generated, as demonstrated in Figure 3.6. The dataset statistics for the first and second splits are summarized in Tables 3.2 and 3.3. Dataset 1 has a rate of the diver_pointing class above 30% for each set: training, validation, and test. Dataset 2 has a slightly lower rate of the diver_pointing class, with a 27.5% occurrence rate.

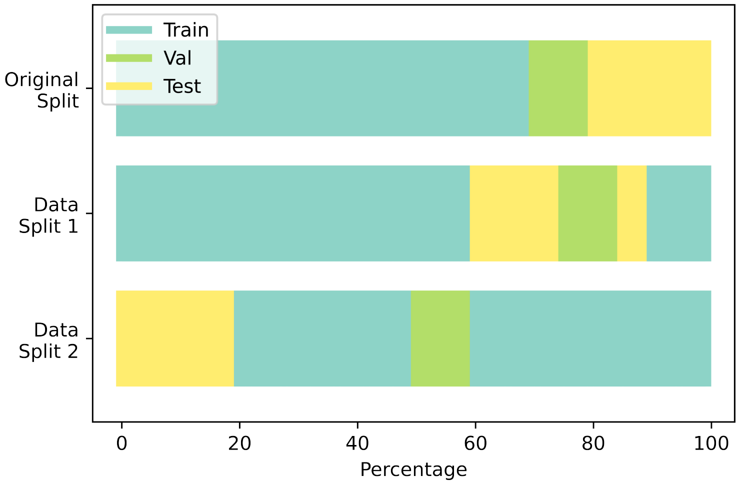


Figure 3.6: Training, validation, and testing dataset split.

Split	# Frames	diver	diver_pointing	% diver	% diver_pointing
train	7899	6076	3258	65.1%	34.9%
val	1131	802	504	61.4%	38.6%
test	2258	1869	851	68.7%	31.3%
total	11288	8747	4613	65.5%	34.5%

Table 3.2: Dataset statistics for Data Split 1.

Split	# Frames	diver	diver_pointing	% diver	% diver_pointing
train	7890	5948	3395	63.7%	36.3%
val	1131	902	497	64.5%	35.5%
test	2267	1897	721	72.5%	27.5%
total	11288	8747	4613	65.5%	34.5%

Table 3.3: Dataset statistics for Data Split 2.

3.3 Deep Learning Algorithms for Pointing Gesture Recognition

To analyze the efficacy of Deep Learning for gesture recognition in the underwater domain, four state-of-the-art (SOTA) object detection networks were trained using the pool trial dataset discussed in §3.2. These include SSD [71] , YOLOv3 [72], YOLOv5 [73], and Faster R-CNN [74], which are each subsequently discussed in more detail.

SSD

The first network we implemented was a Single-Shot Detector (SSD), introduced by Liu *et al.* in [71]. As the name suggests, SSD is a single-stage detector, which identifies regions of interest and bounding boxes for detections simultaneously. We chose to implement a Pytorch-based version of SSD, available on Github [75]. This model uses VGG-16 [76] as the feature extraction backbone to the network, and offered built-in support for datasets in the PASCAL-VOC format. Thus, after setting up our dataset in the required format, SSD was implemented with no major modifications beyond adjustment of the model learning rate.

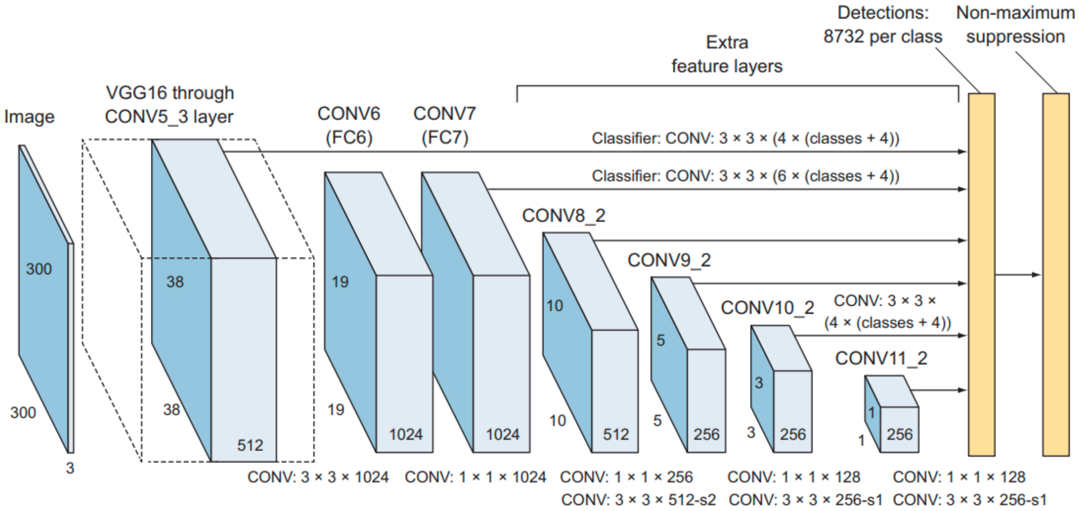


Figure 3.7: SSD Architecture. CONV represents a convolutional layer while FC represents a fully-connected layer. Image courtesy of [6] .

Faster R-CNN

For our second object detection network we shifted focus to the SOTA multi-stage detectors, which first identify regions of interest with associated objective scores, then output final bounding box predictions in a second stage. Currently Faster R-CNN is the state of the art in the family of Region Convolutional Neural Networks (R-CNNs) [74]. Today, Faster R-CNN serves as the backbone for numerous deep learning projects, such as Facebook AI's detectron2 [77], which served as the basis for our implementation of Faster R-CNN.

Implementing Faster R-CNN from the codebase established by detectron2 required two main steps, registration of the dataset and writing custom training and testing scripts. Registration of the dataset has two steps: writing a dataset function which returns the dataset in a standard form compatible with detectron2's existing dataloaders. After the dataset function is written, the dataset is "registered" by entering the name of the dataset along with the dataset function into detectron2's DatasetCatalog. After the dataset name and function have been registered, we provided some optional metadata information, namely the object classes in our dataset, "diver" and "diver_pointing." With the dataset registration complete, the detectron2 codebase is now able to access our custom dataset during training.

The next step in implementing Faster R-CNN involved writing a custom training and evaluation script. This was modified from detectron2's base code and involved writing functions for loading the model configuration, training using this configuration, and running inference on the validation and test sets. The model configuration function is run every time the script is called and includes specification of the base model configuration file, the training dataset, initial model weights, and training parameters such as the batch size, learning rate, number of iterations in an epoch, and number of classes. After the configuration function, the next main function is the core training function, which takes two optional parameters of number of epochs and a resume training flag. This function will then train on the dataset with the established configuration settings, for the specified number of epochs, which defaults to 1. If the resume flag is set to True, training will resume from the default model save checkpoint; if this does not exist, it will begin training from the base checkpoint specified in the configuration. Lastly, two functions were written to run inference: one for the validation set and another for the test set. Each of these functions load the most recently saved checkpoint of the model and run inference on the validation or test set, respectively, saving the output to appropriately named directories. This summarizes the training and evaluation script. By calling this script with the appropriate flags, we can perform any combination of the following tasks: train from the baseline model, resume training from a checkpoint, run validation on the most recent checkpoint, or run test on the most recent checkpoint.

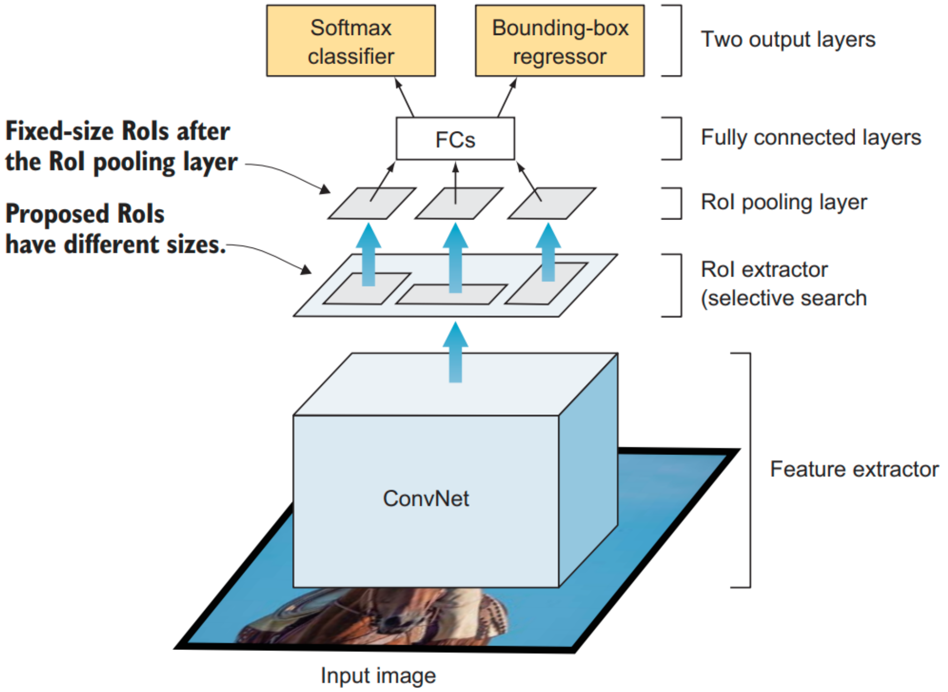


Figure 3.8: Faster R-CNN architecture. ConvNet represents the base feature extractor, while FCs are the fully connected layers. RoI stands for Region of Interest. Image courtesy of [6] .

YOLOv3

YOLOv3 (You Only Look Once) is a SOTA object detection network designed specifically for real-time applications [72]. The YOLO family of networks is uniquely positioned among other object detectors in that it skips the region proposal step and instead makes predictions based on a grid of cells. This modification greatly enhances the speed of the network, with the tradeoff being an increased potential for missed detections. For our implementation, we again train YOLOv3 on our custom dataset, utilizing a pytorch implementation created by Ultralytics [78]. This implementation contained native support for training custom datasets, so model implementation consisted of setting up our dataset in the format required, and writing a YAML configuration file specifying the paths to the training, validation, and test sets, as well as identifying the number of classes and the class names. This step was equivalent to the dataset registration performed for the Faster R-CNN network, with a simplified interface. Once the configuration file was created, it was specified in the built-in training and testing scripts to perform the experiments.

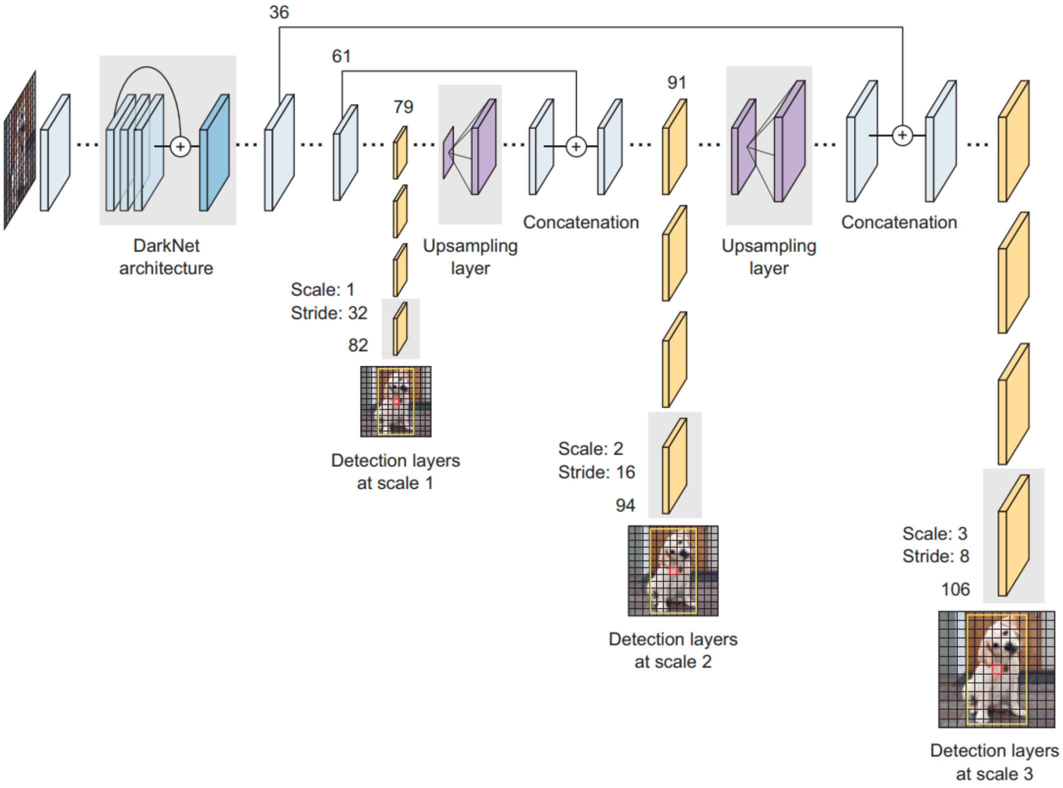


Figure 3.9: YOLOv3 Architecture, Image courtesy of [6] .

YOLOv5

The final implemented network, YOLOv5, is a recently released model which builds upon the YOLO family of networks. Although the maintainers have yet to publish any work on [73], we include it here for comparison as a recent release among the SOTA networks. Also created by Ultralytics, this implementation closely resembles YOLOv3 in structure, thus the details of the implementation setup are equivalent to what is discussed above.

Chapter 4

Experiments

4.1 Pointing Gesture Recognition Experiments

Network Training

When training deep networks based on SOTA models, it is beneficial to begin training with a baseline set of weights from a pretrained model. In this way we can benefit from the general features already learned from the model, training with our new dataset until the model converges. Each of the object detection networks described in §3.3 was trained from a baseline model checkpoint, with the details summarized in Table 4.1. Further details regarding these design choices and the training parameters follow.

SSD. The first model we trained is the Single-Shot Detector (SSD) as implemented by [75]. This model came with a pretrained checkpoint for the base VGG-16 feature

Model	**Feature Extraction Backbone**	**Baseline Model Checkpoint**
SSD	VGG-16	vgg16_reducedfc.pth
Faster R-CNN	ResNet-50	faster_rcnn_R_50_C4_3x.yaml
YOLOv3	N/A	yolov3.pt
YOLOv5	N/A	yolov5x.pt

Table 4.1: Model Feature Extraction Backbones and Baseline Training Checkpoints.

extraction backbone, and from this we trained for 15 epochs with a batch size of 32 and a learning rate of 1e−4, with otherwise default parameters. We observed that the model appeared to converge after 10 epochs as demonstrated in Figure 4.1(d), and based on the model performance, subsequently discussed, we chose the weights after 10 epochs as our final model.

Faster R-CNN. For Faster R-CNN, we chose a baseline model from detectron2's model zoo [77]. Specifically, we chose a model based on the ResNet-50 feature extraction backbone, with a conv4 backbone and conv5 head, which was the original baseline for the Faster R-CNN paper [74]. The particular checkpoint was pre-trained on the COCO dataset [79] utilizing a 3x schedule, which corresponds to approximately 37 epochs of training on the Microsoft Common Objects in Context (COCO) dataset [79]. From this baseline, we trained for 15 epochs (approximately 59k iterations) using a batch size of 2 and a learning rate of 2.5e−4. As shown in Figure 4.1(d), the training converged around 10 epochs, which we took to be our final model to avoid overfitting. Comparison of the metrics discussed below supported this choice, as the average precision (AP) metrics decreased from 10 to 15 epochs.

YOLOv3. The YOLO family of networks contains models of various size, including "tiny" models which slim down the number of parameters in favor of inference speed, while sacrificing model accuracy, and larger models which increase the model parameters for higher accuracy at the cost of inference speed. For an even comparison across implementations, we chose the standard YOLOv3 model, with a checkpoint which was trained to 300 epochs on the COCO dataset. From this checkpoint, we trained on our own datasets for 60 epochs, with an image size of 416 and a batch size of 16. All of the other training parameters were set to the network defaults. With this configuration, we found that the training loss converged after 40 epochs, as demonstrated in Figure 4.1(a). The best checkpoints from training up until 40 epochs and up until 60 epochs were compared for verification, and ultimately we found that the network had generalized better after 40 epochs, suffering from over-fitting past this point.

YOLOv5. The new release of YOLOv5 has a slightly different set of model configurations than YOLOv3. For our baseline we chose the YOLOv5x model, whose checkpoint which achieves the best performance in terms of AP metrics while staying under

100 million parameters. This specific model has 87.7 million parameters in comparison to YOLOv3's 61.9 parameters, and seemed to be the most comparable baseline network in terms of performance and size. The checkpoint we trained from was also trained to 300 epochs on the COCO dataset, providing an even comparison with YOLOv3. For training with our custom dataset, we again used the same parameters as in YOLOv3: training for 60 epochs with an image size of 416 and a batch size of 16, with defaults otherwise. Very similarly to YOLOv3, we found that our network converged after 40 epochs, and we took our final model to be the best checkpoint from the initial 40 epochs.

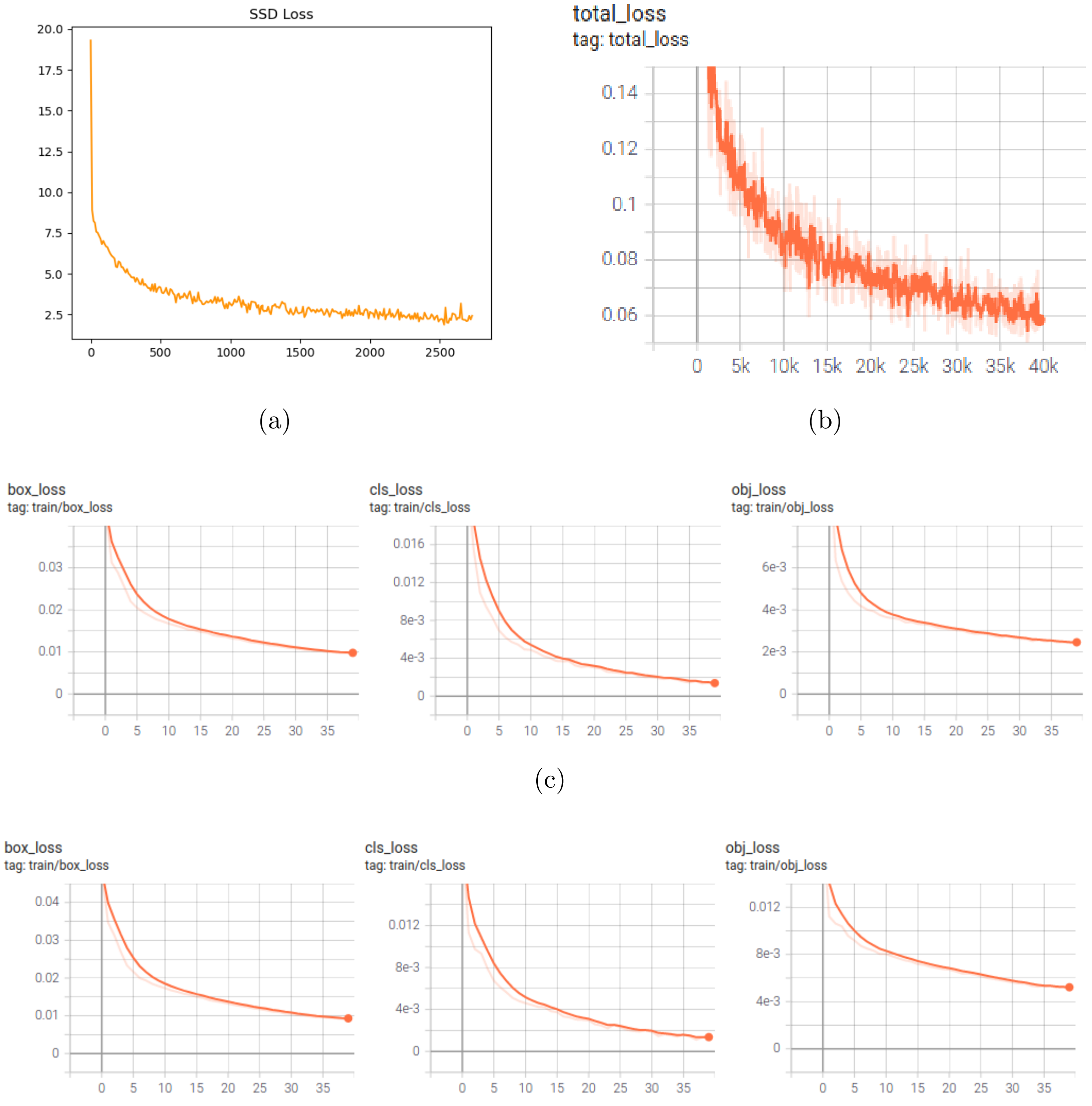
Network Performance Comparison

Metrics

Here we present three key metrics to analyze the performance of our object detection networks in recognizing pointing gestures: Frames per Second (FPS), Intersection over Union (IOU) and Average Precision (AP).

FPS. Frames per second (FPS) is the standard metric for measuring detection speed of an object detector, and refers to the number of frames the network can make predictions for each second, *i.e.*, the number of forward passes the network can make in one second [6]. As shown in Table 4.2, YOLOv3 has the best inference speed, at a minimum of 129 FPS between the two datasets. YOLOv5 has a slightly different architecture and more weights, which likely contributes to its decreased framerate at 83 FPS. Since the YOLO family of object detection networks were build specifically for real-time inference and eliminate the region proposal step altogether, their speed comes as no surprise. After the YOLO detectors, SSD comes in next, with a drop in inference speed to 38 FPS, roughly half of the speed of YOLOv5. This is to be expected, with the VGG-16 feature extraction adding additional overhead in this single-shot detector. Despite the speed drop in comparison to YOLOv5, this is still fast enough for real-time inference on board an AUV. Lastly, Faster R-CNN sees another significant drop in inference speed, down to 5 FPS due to the multi-stage architecture. Dropping below 15 FPS, this is no longer considered real-time; however, this is still fast enough to be used on board an AUV.

IOU. The Intersection over Union (IOU) metric quantifies the overlap between the



(a)

(b)

(c)

(d)

Figure 4.1: Training loss of each of the networks (a) SSD, (b) Faster R-CNN as implemented in Detectron2, (c) YOLOv3, and (d) YOLOv5. Loss is plotted on the y-axis and number of epochs (YOLOv3, YOLOv5) or number of iterations (SSD, Faster R-CNN) is on the x-axis.

	Model	**inference speed**	**FPS**
Data Split 1	SSD	26.1 ms	38.3
	Faster R-CNN	198 ms	5.1
	Yolov3	7.6 ms	131.6
	Yolov5	12.0 ms	83.3
Data Split 2	SSD	24.9 ms	40.1
	Faster R-CNN	208.1 ms	4.8
	Yolov3	7.7 ms	129.9
	Yolov5	10.4 ms	96.2

Table 4.2: Inference speed on NVIDIA Titan Xp GPU.

original ground truth bounding box and the bounding box predicted by the object detector. A value between zero and one, it is calculated by a dividing the area of intersection between the ground truth bounding box and the predicted bounding box by the area of their union. Given BB_{gt} as the ground truth and BB_{pred} as the predicted bounding box, this can be expressed mathematically as

$$IOU = \frac{BB_{gt} \cap BB_{pred}}{BB_{gt} \cup BB_{pred}}. \tag{4.1}$$

The closer the IOU value is to one, the higher the overlap between the ground truth and predicted bounding boxes, a reflection on the spatial accuracy of the object detector. A demonstration of this may be seen in Figure 4.2

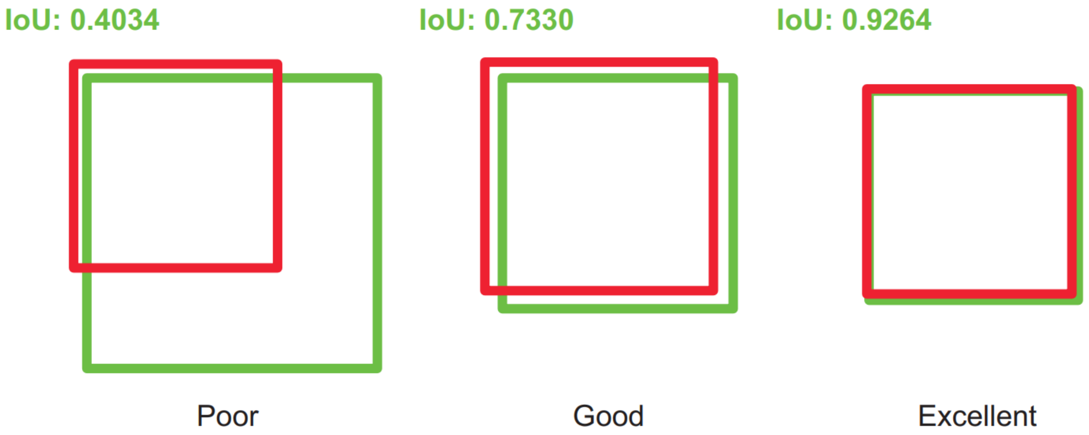


Figure 4.2: Demonstration of IOU scores, Image courtesy of [6]

Beyond being a measure of the spatial accuracy of the object detector, the IOU metric is further used to determine when a detection is considered a True Positive (TP), *i.e.*, a correct prediction. A threshold for the IOU value is set as a network-tunable value, where given a class label, all detections with an IOU between the ground truth and the detection greater than the threshold are considered True Positives, while detections whose IOU is below the threshold are considered False Positives (FP). Missed detections or detections with the incorrect class label are considered False Negatives (FN).

Overall, the four networks analyzed here performed extremely well in terms of average IOU, with all of the networks achieving IOU scores above 83% (Table 4.3). This represents more than 10% improvement over the average IOU scores achieved by SSD and the YOLO family of networks when trained to detect divers using the VDD-$\bar{C}$ and DDD datasets [80], although this improvement may be partially attributed to our dataset lacking the challenging visibility conditions present in these datasets. For both data splits, SSD has the highest IOU scores, 99.3% and 99.4% respectively, indicating that this network has effectively localized our detections. The YOLO family of networks saw an improvement from the 83rd percentile to the 96th percentile between the two data splits, indicating the strong potential of the network architecture to learn to localize effectively. Faster R-CNN held to steady IOU values around 87%. These results suggest that our detectors were able to effectively learn the features of a diver and diver_pointing for accurate localization.

AP. Lastly, one of the key benchmark metrics for object detectors is the Average Precision (AP). This metric is calculated from the precision and recall, which are defined based on the True Positives (TP), False Positives (FP), and False Negatives (FN) as

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$recall = \frac{TP}{TP + FN}. \tag{4.3}$$

For a set IOU threshold and a single class, the precision and recall can be computed over a range of bounding box confidence score thresholds to create what is known as the precision-recall curve. The average precision (AP) is then computed as the area under this curve. This process is repeated for each class, and the mean of the average precision values taken over all the classes defines the mean average precision (mAP). Specifically when referring to the COCO dataset, AP is used to refer to the mAP, and we adopt this convention in subsequent discussions here. Further, since the AP is dependent on the IOU threshold, it is conventional to specify the threshold along with the AP; thus the mAP calculated with an IOU threshold of 0.5 is denoted AP_{50}, while the mAP with an IOU threshold of 0.75 is denoted AP_{75}. One final metric, associated specifically with the COCO dataset, is an average AP over IOU thresholds. This metric, $AP_{50:95}$ calculates the APs with thresholds between 0.5 and 0.95 inclusive, with a step size of 0.05, and computes their average.

Analyzing the AP values reveals that for both data splits, the YOLO family of object detectors has superior performance, with YOLOv5 being the most effective detector for data split 1, and YOLOv3 being marginally superior for data split 2. Considering the AP_{50} metric specifically, the models trained on data split 1 have comparable performance, with a narrow range of values between 82% and 90%. Data split 2 revealed some interesting trends in the ability of the models to learn, with SSD and Faster R-CNN's AP_{50} metrics dropping from 82% to 65.9%, and 87.1% to 80.9% respectively, while YOLOv3 and YOLOv5 both increased to an AP_{50} of 99.7%. We attribute the increase in the YOLO models' performance to their increased recall. Looking at Table 4.4, the number of frames containing a diver or diver_pointing instance which had no

		diver	**diver_pointing**	**Both Classes**			
	Model	**AP$_{.50:.95}$**	**AP$_{.50:.95}$**	**AP$_{50}$**	**AP$_{75}$**	**AP$_{.50:.95}$**	**IOU**
Data Split 1	SSD	0.446	0.542	0.821	0.568	0.494	**0.993**
	Faster R-CNN	0.598	0.735	0.871	0.738	0.666	0.870
	Yolov3	0.769	0.782	0.876	-	0.775	0.833
	Yolov5	**0.775**	**0.845**	**0.9**	-	**0.81**	0.837
Data Split 2	SSD	0.486	0.353	0.659	0.506	0.420	**0.994**
	Faster R-CNN	0.669	0.587	0.809	0.696	0.628	0.890
	Yolov3	**0.998**	**0.996**	**0.997**	-	**0.997**	0.968
	Yolov5	0.997	0.996	0.997	-	0.997	0.968

Table 4.3: Summary of AP and IOU Metrics across the two data splits and four object detectors.

corresponding prediction dropped from a rate of close to 8% on the first data split, down to a rate of 0.5%. These frames with no predictions (*i.e.*, missed detections) are counted as false negatives, so this decrease contributes significantly to improved recall. Considering the $AP_{50:95}$ metric across both data splits, YOLO remains strong, with averaging over increasing IOU thresholds not diminishing model performance significantly.

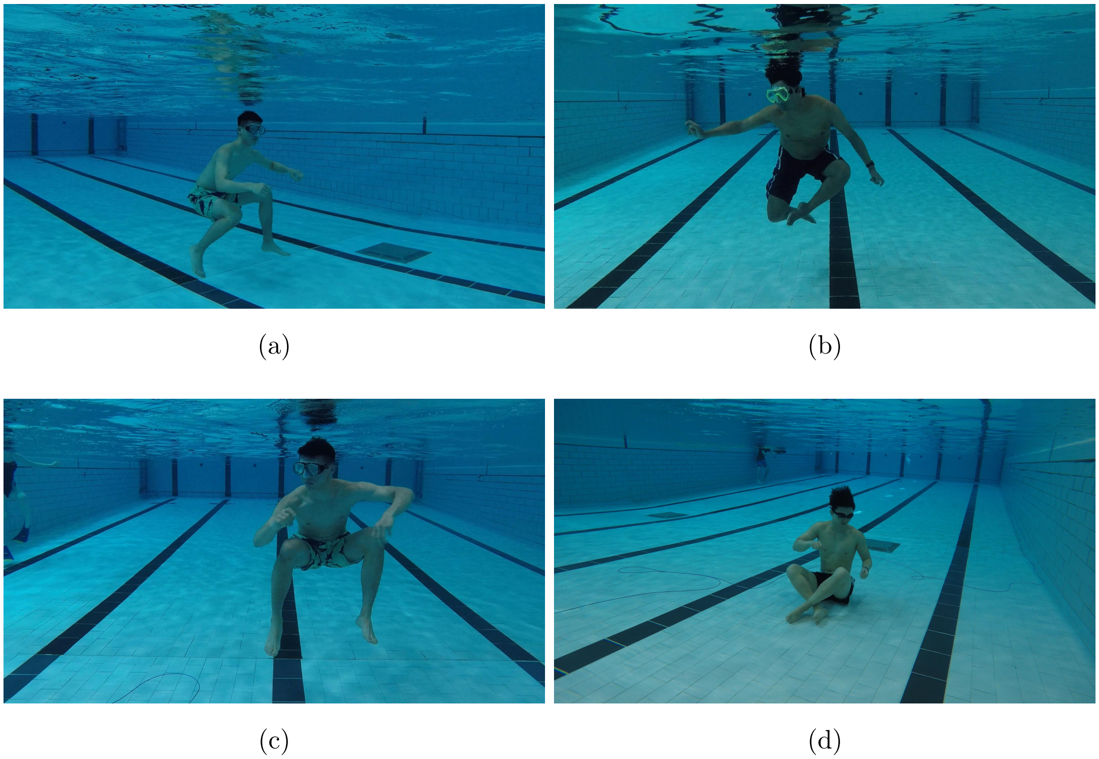
Finally, we performed a more qualitative analysis of some of the classification failure modes. Several of the misclassifications for data split 1 with YOLOv3 may be seen in Figures 4.3 and 4.4. Figure 4.3 shows four frames where the ground truth prediction was diver, but the network prediction was diver_pointing. One commonality which is demonstrated in Figures 4.3a and 4.3b is the outstretched arm, which appears to indicate that the network has learned to correlate an extended arm to a pointing gesture, which aligns with one of the rules in our labeling policy. Further, both Figure 4.3b and 4.3c have the index finger partially extended from the rest of the hand and appear to be frames just prior to a sequence of pointing gestures, supporting the idea that our network has learned a representation of a pointing hand pose and associated that with the label diver_pointing. In Figure 4.4 we consider the frames with a ground truth

		Missed Detections			Rate
	Model	**diver**	**diver_pointing**	**total**	**all classes**
Data Split 1	SSD	0	0	0	0%
	Faster R-CNN	59	0	59	2.2%
	Yolov3	197	22	219	8.1%
	Yolov5	195	21	216	7.9%
Data Split 2	SSD	0	0	0	0%
	Faster R-CNN	17	0	17	0.6%
	Yolov3	3	9	12	0.5%
	Yolov5	3	9	12	0.5%

Table 4.4: Missed detections summary by dataset and object detector.

label of diver_pointing which were misclassified as diver. In Figure 4.4a, the diver's arm is not fully extended, and further, the pointing finger is difficult to distinguish, being a probable cause for the misclassification. Figures 4.4b and 4.4c demonstrate failure modes when the divers' hands are in the full-hand pointing pose. We surmise that our network may not have learned a good representation for the full-hand point, or that it has not associated this representation with the pointing class. Our last failure mode demonstrated here is in Figure 4.4d, where the diver's extended arm aligns with the body while making a pointing gesture. This implies that the network has learned to associate pointing gestures with an arm pose extended away from the body, thus missing this pointing gesture. Each of these failure modes gives insight into what our network has learned and demonstrate the challenging edge cases that future models can be developed to more robustly handle.

After an analysis of these four object detection frameworks, we believe that YOLOv3 demonstrates the overall best performance in terms of AP, with the added advantage of a framerate sufficiently high enough for real-time object detection. This high framerate does compensate partially for the missed detection rate that manifested itself for one of the data splits.



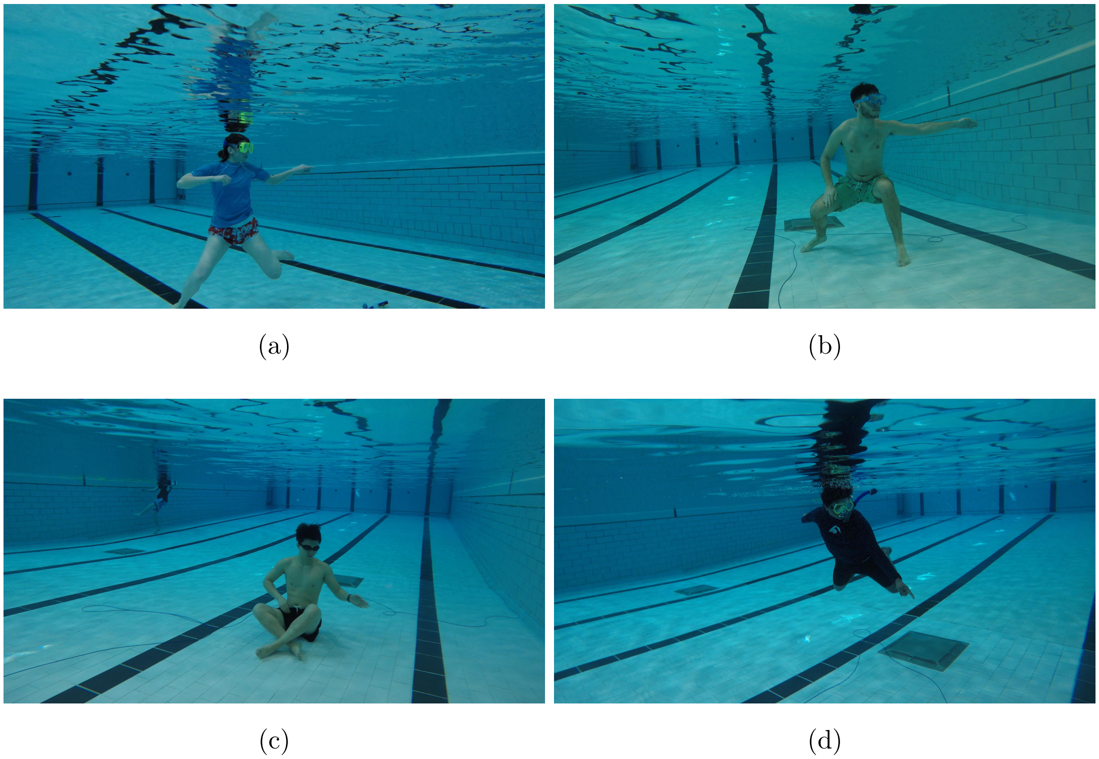
(a)

(b)

(c)

(d)

Figure 4.3: Images with ground truth labels of “diver” which were misclassified as “diver_pointing.”



(a)

(b)

(c)

(d)

Figure 4.4: Images with ground truth labels of "diver_pointing" which were misclassified as "diver."

Chapter 5

Conclusion and Future Work

A robust, reliable perception algorithm for the recognition of pointing gestures is an important stepping stone towards effective human-robot collaboration with autonomous underwater vehicles. Since pointing gestures provide a natural interaction interface for divers in a demanding aquatic environment, we choose them as a fundamental communication gesture and lay the groundwork for implementing a pointing gesture recognition algorithm. In contrast to the abundance of work in identifying pointing gestures in the terrestrial domain, we present here the first work exploring pointing gesture recognition underwater.

While focusing on pointing gestures for underwater robotics, we perform a human study to investigate the various ways in which individuals perform our four proposed sub-classes of pointing gestures in a terrestrial environment, with the intent of gathering sufficient data to train a model to both identify these as pointing gestures, and ultimately differentiate between them. This model would then serve as a baseline pretrained model which could be adapted to the underwater environment once domain data has been collected.

Further, we contribute two annotated datasets for pointing gestures, one in the aquatic and a second in the terrestrial domain, focused on the identification of pointing gestures from still frames, and annotated for training object detectors. The first dataset was subsequently used to train four state-of-the-art object detectors, to compare their efficacy for detecting pointing gestures based on a single frame.

Our contributions lay the groundwork for investigating pointing gestures in underwater human-robot interaction, for the particular purpose of collaborating with autonomous underwater vehicles. A natural extension of this work is the differentiation between the four classes of pointing gestures presented in our human study, for the purposes of deriving diver intent from gesture form; this topic is currently being investigated by a fellow member of the Interactive Robotics and Vision Lab, Luoyao Chen. Still focusing on the pointing gesture recognition problem, further approaches for robustly identifying a pointing gesture could be considered, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) models which take into consideration a temporal sequence of frames rather than a single static frame. Beyond identification of the pointing gestures and their intent, there is further exploration to be done in the area of identifying the object or direction indicated by a pointing gesture. In sparse aquatic environments, deep learning algorithms leveraging object detection present an opportunity for inference of the object indicated by a pointing gesture; such an algorithm could be inspired by the methodologies of Wang *et al.* in [81], where a human-object interaction is learned by a deep network. A more fundamental approach could infer the pointing gesture direction from diver pose estimation, much like the terrestrial techniques leveraging the Kinect sensor have implemented.

This thesis presents novel work towards developing a robust algorithm for recognizing pointing gestures underwater, and lays a framework for future research in developing natural, effective, and intuitive human-robot interaction with autonomous underwater vehicles.

Bibliography

[1] Gregory Dudek, Philippe Giguere, Chris Prahacs, Shane Saunderson, Junaed Sattar, Luz-abril Torres-Mendez, Michael Jenkin, Andrew German, Andrew Hogue, Arlene Ripsman, Jim Zacher, Evangelos Milios, Hui Liu, Pifu Zhang, Marti Buehler, and Christina Georgiades. AQUA: An Amphibious Autonomous Robot. *Computer*, 40(1):46–53, January 2007.

[2] Chelsey Edge, Sadman Sakib Enan, Michael Fulton, Jungseok Hong, Jiawei Mo, Kimberly Barthelemy, Hunter Bashaw, Berik Kallevig, Corey Knutson, Kevin Orpen, and Junaed Sattar. Design and Experiments with LoCO AUV: A Low Cost Open-Source Autonomous Underwater Vehicle. *CoRR*, 2020.

[3] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing Underwater Imagery Using Generative Adversarial Networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7159–7165, 2018.

[4] Anqi Xu, Gregory Dudek, and Junaed Sattar. A Natural Gesture Interface for Operating Robotic Systems. In *2008 IEEE International Conference on Robotics and Automation*, pages 3557–3563, May 2008. ISSN: 1050-4729.

[5] Md Jahidul Islam, Marc Ho, and Junaed Sattar. Dynamic Reconfiguration of Mission Parameters in Underwater Human-Robot Collaboration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6212–6219, May 2018. ISSN: 2577-087X.

[6] Mohamed Elgendy. *Deep Learning for Vision Systems*. Simon and Schuster, November 2020. Google-Books-ID: 97YCEAAAQBAJ.

[7] Tony Greicius. Mars Perseverance Rover. http://www.nasa.gov/perseverance, 2016. [Online; accessed May-19-2021].

[8] Ocean | Definition, Distribution, Map, Formation, & Facts. https://www.britannica.com/science/ocean, 2021. [Online; accessed May-19-2021].

[9] Recreational Scuba Training Council. Common hand signals for recreational scuba diving. *Online pdf Available http://www.neadc.org/CommonHandSignalsforScubaDiving.pdf.[Accessed Sept. 20 2019]*, 2014.

[10] Sotaro Kita. *Pointing: Where Language, Culture, and Cognition Meet.* Psychology Press, June 2003. Google-Books-ID: JlN4AgAAQBAJ.

[11] National Geographic Society. Great Pacific Garbage Patch. http://www.nationalgeographic.org/encyclopedia/great-pacific-garbage-patch/, July 2019. [Online; accessed May-19-2021].

[12] Michael Fulton, Jungseok Hong, Md Jahidul Islam, and Junaed Sattar. Robotic Detection of Marine Litter Using Deep Visual Detection Models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5752–5758, May 2019. ISSN: 2577-087X.

[13] Underwater Weapons. https://fas.org/man/dod-101/navy/docs/es310/uw_wpns/uw_wpns.htm. [Online; accessed May-19-2021].

[14] UDMessenger - Underwater Robotic Device. http://www1.udel.edu/udmessenger/vol19no1/stories/research_robotic-device.html. [Online; accessed May-19-2021].

[15] Barbara Martinez, Jamie K. Reaser, Alex Dehgan, Brad Zamft, David Baisch, Colin McCormick, Anthony J. Giordano, Rebecca Aicher, and Shah Selbe. Technology Innovation: Advancing Capacities for the Early Detection of and Rapid Response to Invasive Species. *Biological Invasions*, 22(1):75–100, January 2020.

[16] Christian A. Mueller, Tobias Fromm, Heiko Buelow, Andreas Birk, Maximilian Garsch, and Norbert Gebbeken. Robotic Bridge Inspection within Strategic Flood Evacuation Planning. In *OCEANS 2017 - Aberdeen*, pages 1–6, June 2017.

[17] Arturo Gomez Chavez, Christian A. Mueller, Tobias Doernbach, Davide Chiarella, and Andreas Birk. Robust Gesture-Based Communication for Underwater Human-Robot Interaction in the Context of Search and Rescue Diver Missions. *Journal of Marine Science and Engineering*, 7(1):16, January 2019.

[18] Kensei Ishizu, Norimitsu Sakagami, Kouhei Ishimaru, Mizuho Shibata, Hiroyuki Onishi, Shigeo Murakami, and Sadao Kawamura. Ship Hull Inspection Using a Small Underwater Robot with a Mechanical Contact Mechanism. In *2012 Oceans - Yeosu*, pages 1–6, May 2012.

[19] A. Vasilijevic, Ð Nađ, N. Stilinovic, N. Mišković, and Z. Vukic. Application of an ASV for Coastal Underwater Archaeology. 2016.

[20] Arturo Gomez Chavez, Jorge Fontes, Pedro Afonso, Max Pfingsthorn, and Andreas Birk. Automated Species Counting Using a Hierarchical Classification Approach with Haar Cascades and Multi-descriptor Random Forests. In *OCEANS 2016 - Shanghai*, pages 1–6, April 2016.

[21] Derya Akkaynak and Tali Treibitz. A Revised Underwater Image Formation Model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6723–6732, June 2018. ISSN: 2575-7075.

[22] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer. Audiovisual Person Tracking with a Mobile Robot. In *In Proc. Int. Conf. on Intelligent Autonomous Systems*, pages 898–906. IOS Press, 2004.

[23] Jacques Penders and Ayan Ghosh. Human Robot Interaction in the Absence of Visual and Aural Feedback: Exploring the Haptic Sense. *Procedia Computer Science*, 71:185–195, January 2015.

[24] Chung Hyuk Park and Ayanna M. Howard. Towards Real-time Haptic Exploration Using a Mobile Robot as Mediator. In *2010 IEEE Haptics Symposium*, pages 289–292, March 2010. ISSN: 2324-7355.

[25] Mirko DiGiacomcantonio and Yonathan Gebreyes. Self-propelled Luggage. `https://patents.google.com/patent/US20140107868A1/en`, April 2014.

[26] Guillaume Doisy, Aleksandar Jevtić, and Saša Bodiroža. Spatially Unconstrained, Gesture-Based Human-Robot Interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 117–118, March 2013. ISSN: 2167-2148.

[27] Hugh Durrant-Whyte, Nicholas Roy, and Pieter Abbeel. Comparing Heads-Up, Hands-Free Operation of Ground Robots to Teleoperation. In *Robotics: Science and Systems VII*, pages 193–200. MIT Press, 2012.

[28] Sushmita Mitra and Tinku Acharya. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, May 2007.

[29] Gerard Canal, Sergio Escalera, and Cecilio Angulo. A Real-Time Human-Robot Interaction System Based on Gestures for Assistive Scenarios. *Computer Vision and Image Understanding*, 149:65–77, August 2016.

[30] Alex Couture-Beil, Richard T. Vaughan, and Greg Mori. Selecting and Commanding Individual Robots in a Vision-Based Multi-Robot System. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 355–356, March 2010. ISSN: 2167-2148.

[31] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, February 1978.

[32] Calin Belta, Antonio Bicchi, Magnus Egerstedt, Emilio Frazzoli, Eric Klavins, and George J. Pappas. Symbolic Planning and Control of Robot Motion [Grand Challenges of Robotics]. *IEEE Robotics Automation Magazine*, 14(1):61–70, March 2007.

[33] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with Trust for Human-Robot Collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI ’18, pages 307–315, New York, NY, USA, February 2018. Association for Computing Machinery.

[34] Kevin J. DeMarco, Michael E. West, and Ayanna M. Howard. Underwater Human-Robot Communication: A Case Study with Human Divers. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3738–3743, October 2014. ISSN: 1062-922X.

[35] A.D. Wilson and A.F. Bobick. Parametric Hidden Markov Models for Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, September 1999.

[36] Kai Nickel and Rainer Stiefelhagen. Pointing Gesture Recognition Based on 3D-Tracking of Face, Hands and Head Orientation. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, ICMI '03, pages 140–146, New York, NY, USA, November 2003. Association for Computing Machinery.

[37] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A Multimedia Gesture Dataset for Human Robot Communication: Acquisition, Tools and Recognition Results. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3066–3070, September 2016. ISSN: 2381-8549.

[38] ChaLearn. `http://chalearnlap.cvc.uab.es/`.

[39] Farhood Negin, Pau Rodriguez, Michal Koperski, Adlen Kerboua, Jordi Gonzàlez, Jeremy Bourgeois, Emmanuelle Chapoulie, Philippe Robert, and Francois Bremond. PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition. *Expert Systems with Applications*, 106:21–35, 2018.

[40] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A Low Power, Fully Event-Based Gesture Recognition System. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, Honolulu, HI, July 2017. IEEE.

[41] Simon Ruffieux, Denis Lalanne, Elena Mugellini, and Omar Abou Khaled. A Survey of Datasets for Human Gesture Recognition. In Masaaki Kurosu, editor, *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, Lecture

Notes in Computer Science, pages 337–348, Cham, 2014. Springer International Publishing.

[42] Christopher Conly, Paul Doliotis, Pat Jangyodsuk, Rommel Alonzo, and Vassilis Athitsos. Toward a 3D Body Part Detection Video Dataset and Hand Tracking Benchmark. In *Proceedings of the 6th International Conference on Pervasive Technologies Related to Assistive Environments*, PETRA '13, pages 1–6, New York, NY, USA, May 2013. Association for Computing Machinery.

[43] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing Actions by Shape-Motion Prototype Trees. In *2009 IEEE 12th International Conference on Computer Vision*, pages 444–451, September 2009. ISSN: 2380-7504.

[44] Sergio Escalera, Cristian Sminchisescu, Richard Bowden, Stan Sclaroff, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Isabelle Guyon, Vassilis Athitsos, Hugo Escalante, Leonid Sigal, and Antonis Argyros. ChaLearn Multi-modal Gesture Recognition 2013: Grand Challenge and Workshop Summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 365–368, Sydney, Australia, 2013. ACM Press.

[45] Dai Fujita and Takashi Komuro. 3D Pose Estimation of a Front-Pointing Hand Using a Random Regression Forest. In Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, editors, *Computer Vision – ACCV 2016 Workshops*, Lecture Notes in Computer Science, pages 197–211, Cham, 2017. Springer International Publishing.

[46] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic Detection of Pointing Directions for Human-Robot Interaction. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, November 2015.

[47] Yuhui Lai, Chen Wang, Yanan Li, Shuzhi Sam Ge, and Deqing Huang. 3D Pointing Gesture Recognition for Human-Robot Interaction. In *2016 Chinese Control and Decision Conference (CCDC)*, pages 4959–4964, May 2016. ISSN: 1948-9447.

[48] Satoshi Ueno, Sei Naito, and Tsuhan Chen. An Efficient Method for Human Pointing Estimation for Robot Interaction. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1545–1549, October 2014. ISSN: 2381-8549.

[49] Michal Tölgyessy, Martin Dekan, František Duchoň, Jozef Rodina, Peter Hubinský, and L'uboš Chovanec. Foundations of Visual Linear Human–Robot Interaction via Pointing Gesture Navigation. *International Journal of Social Robotics*, 9(4):509–523, September 2017.

[50] Michal Tölgyessy, Martin Dekan, and Peter Hubinský. Human-Robot Interaction Using Pointing Gestures. In *Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control*, ISCSIC '18, pages 1–5, New York, NY, USA, September 2018. Association for Computing Machinery.

[51] Ye-Peng Guan, Hui-Jun Xiong, and Yun-Jie Yu. Single Camera Vision Pointing Recognition for Natural HCI. In *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2009)*, pages 106–108, December 2009.

[52] Ye-Peng Guan. Uncalibrated Camera Vision Pointing Recognition for HCI. In *2010 13th IEEE International Conference on Computational Science and Engineering*, pages 204–207, December 2010.

[53] Maria Pateraki, Haris Baltzakis, and Panos Trahanias. Visual Estimation of Pointed Targets for Robot Guidance Via Fusion of Face Pose and Hand Orientation. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1060–1067, November 2011.

[54] Jan Richarz, Andrea Scheidig, Christian Martin, Steffen Müller, and Horst-Michael Gross. A Monocular Pointing Pose Estimator for Gestural Instruction of a Mobile Robot. *International Journal of Advanced Robotic Systems*, 4(1):17, March 2007. Publisher: SAGE Publications.

[55] Christian Martin, Frank-Florian Steege, and Horst-Michael Gross. Estimation of Pointing Poses for Visually Instructing Mobile Robots Under Real World Conditions. *Robotics and Autonomous Systems*, 58(2):174–185, February 2010.

[56] Gregory Dudek, Junaed Sattar, and Anqi Xu. A Visual Language for Robot Control and Programming: A Human-Interface Study. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2507–2513, April 2007. ISSN: 1050-4729.

[57] M. Fiala. ARTag, a Fiducial Marker System Using Digital Techniques. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 590–596 vol. 2, June 2005. ISSN: 1063-6919.

[58] P.J. Besl and Neil D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992.

[59] D. Chiarella, M. Bibuli, G. Bruzzone, M. Caccia, A. Ranieri, E. Zereik, L. Marconi, and P. Cutugno. Gesture-based Language for Diver-Robot Underwater Interaction. In *OCEANS 2015 - Genova*, pages 1–9, May 2015.

[60] Nikola Mišković, Antonio Pascoal, Marco Bibuli, Massimo Caccia, Jeffrey A. Neasham, Andreas Birk, Murat Egi, Karl Grammer, Alessandro Marroni, Antonio Vasilijević, and Zoran Vukić. CADDY Project, Year 2: The First Validation Trials. *IFAC-PapersOnLine*, 49(23):420–425, January 2016.

[61] Nikola Mišković, Antonio Pascoal, Marco Bibuli, Massimo Caccia, Jeffrey A. Neasham, Andreas Birk, Murat Egi, Karl Grammer, Alessandro Marroni, Antonio Vasilijević, Đula Nađ, and Zoran Vukić. CADDY Project, Year 3: The Final Validation Trials. In *OCEANS 2017 - Aberdeen*, pages 1–5, June 2017.

[62] Davide Chiarella, Marco Bibuli, Gabriele Bruzzone, Massimo Caccia, Andrea Ranieri, Enrica Zereik, Lucia Marconi, and Paola Cutugno. A Novel Gesture-Based Language for Underwater Human–Robot Interaction. *Journal of Marine Science and Engineering*, 6(3):91, September 2018.

[63] Matt Ervin G. Mital, Herbert V. Villaruel, and Elmer P. Dadios. Neural Network Implementation of Divers Sign Language Recognition based on Eight Hu-Moment Parameters. In *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–6, October 2018.

[64] Manuel Zahn. Development of an Underwater Hand Gesture Recognition System. In *Global Oceans 2020: Singapore – U.S. Gulf Coast*, pages 1–8, October 2020. ISSN: 0197-7385.

[65] Đula Nađ, Christopher Walker, Igor Kvasić, Derek Orbaugh Antillon, Nikola Mišković, Iain Anderson, and Ivan Lončar. Towards Advancing Diver-Robot Interaction Capabilities. *IFAC-PapersOnLine*, 52(21):199–204, January 2019.

[66] Franka Gustin, Ivor Rendulic, Nikola Miskovic, and Zoran Vukic. Hand Gesture Recognition from Multibeam Sonar Imagery. *IFAC-PapersOnLine*, 49(23):470–475, January 2016.

[67] Belmont Report. In *Encyclopedia of Public Health*, pages 60–60. Springer Netherlands, Dordrecht, 2008.

[68] B-Roll Video Index - Grand Canyon National Park (U.S. National Park Service). `https://www.nps.gov/grca/learn/photosmultimedia/b-roll_hd_index.htm`. [Online; accessed May-19-2021].

[69] GoPro | World's Most Versatile Cameras | Shop Now & Save. `https://gopro.com/en/us/`. [Online; accessed May-19-2021].

[70] Ericsson/eva. `https://github.com/Ericsson/eva`, October 2020. [Online; accessed May-19-2021].

[71] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. *arXiv:1512.02325 [cs]*, 9905:21–37, 2016.

[72] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, April 2018.

[73] ultralytics/yolov5. `https://github.com/ultralytics/yolov5`, May 2021. [Online; accessed May-19-2021].

[74] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*, January 2016.

[75] Max deGroot. amdegroot/ssd.pytorch. `https://github.com/amdegroot/ssd.pytorch`, April 2021. [Online; accessed May-19-2021].

[76] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015.

[77] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[78] ultralytics/yolov3. `https://github.com/ultralytics/yolov3`, May 2021. [Online; accessed May-19-2021].

[79] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February 2015.

[80] Karin de Langis, Michael Fulton, and Junaed Sattar. An Analysis of Deep Object Detectors For Diver Detection. *arXiv:2012.05701 [cs]*, November 2020.

[81] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning Human-Object Interaction Detection Using Interaction Points. *arXiv:2003.14023 [cs]*, March 2020.

Appendix A

IRB Human Study Details

This Appendix contains supplementary materials relevant to the Human Study discussed in §3.1. First, we provide our Social Protocol document (HRP-580) and the Social-Behavioral Consent Form (HRP-582) in §A.1 and §A.2. Both of these forms were part of our original IRB submission. Next we provide our exemption determination (§A.3) which permitted the replacement of the consent form with the Information Sheet for Research (HRP-587) in §A.4. Our final approval document is reproduced in §A.5, and the promotional materials and intake form for our study are included in §A.6 and §A.7.

A.1 Social Protocol, HRP-580

ANCILLARY REVIEWS

Which ancillary reviews do I need and when do I need them?

Refer to HRP-309 for more information about these ancillary reviews.

Select yes or no	Does your study…	*If yes…*	Impact on IRB Review
☐ Yes ☒ No	Include Gillette resources, staff or locations	*Gillette Scientific review and Gillette Research Administration approval is required. Contact:* *research@gillettechildrens.com*	**Required prior to IRB submission**
☐ Yes ☒ No	Involve Epic, or Fairview patients, staff, locations, or resources?	*The Fairview ancillary review will be assigned to your study by IRB staff* *Contact:* ancillaryreview@Fairview.org	**Approval must be received prior to IRB committee / designated review.** **Consider seeking approval prior to IRB submission.**
☐ Yes ☒ No	Include evaluation of drugs, devices, biologics, tobacco, or dietary supplements or data subject to FDA inspection?	*STOP – Complete the Medical Template Protocol (HRP-590)* *The regulatory ancillary review will be assigned to your study by IRB staff* *Contact: medreg@umn.edu* *See https://policy.umn.edu/research/indide*	
☐ Yes ☒ No	Require Scientific Review? Not sure? See guidance on next page.	*ONLY REQURED BIOMEDICAL RESEARCH REVIEWED BY FULL COMMITTEE*	
☐ Yes ☒ No	Relate to cancer patients, cancer treatments, cancer screening/prevention, or tobacco?	*Complete the CPRC application process.* *Contact: ccprc@umn.edu*	
☐ Yes ☒ No	Include the use of radiation? (x-ray imaging, radiopharmaceuticals, external beam or brachytherapy)	*Complete the AURPC Human Use Application and follow instructions on the form for submission to the AURPC committee.* *Contact: barmstro@umn.edu*	**Approval from these committees must be received prior to IRB approval;**

☐ **Yes** ☒ **No**	Use the Center for Magnetic Resonance Research (CMRR) as a study location?	*Complete the [CMRR pre-IRB ancillary review]* *Contact: ande2445@umn.edu*	**These groups each have their own application process.**
☐ **Yes** ☒ **No**	Include the use of recombinant or synthetic nucleic acids, toxins, or infectious agents?	*STOP – Complete [the Medical Template Protocol (HRP-590)]*	
☐ **Yes** ☒ **No**	Include the use of human fetal tissue, human embryos, or embryonic stem cells?	*STOP – Complete [the Medical Template Protocol (HRP-590)]*	
☐ **Yes** ☒ **No**	Include PHI or are you requesting a HIPAA waiver?	*If yes, HIPCO will conduct a review of this protocol.* *Contact: privacy@umn.edu*	
☐ **Yes** ☒ **No**	Use data from the Information Exchange (IE)?	*The Information Exchange ancillary review will be assigned to your study by IRB staff* *Contact: ics@umn.edu*	**Approval must be received prior to IRB approval.** **These groups do not have a separate application process but additional information from the study team may be required.**
☐ **Yes** ☒ **No**	Use the Biorepository and Laboratory Services to collect tissue for research?	*STOP – Complete [the Medical Template Protocol (HRP-590)]* *The BLS ancillary review will be assigned to your study by IRB staff.* *Contact: cdrifka@umn.edu*	
☐ **Yes** ☒ **No**	Have a PI or study team member with a conflict of interest?	*The CoI ancillary review will be assigned to your study by IRB staff* *Contact: becca002@umn.edu*	
☐ **Yes** ☒ **No**	Need to be registered on clinicaltrials.gov?	*If you select "No" in ETHOS, the clinicaltrials.gov ancillary review will be assigned to your study by IRB staff* *Contact: kmmccorm@umn.edu*	
☐ **Yes** ☒ **No**	Require registration in OnCore?	*If you select "No" or "I Don't Know" in ETHOS, the OnCore ancillary review will be assigned to your study by IRB staff* *Contact: oncore@umn.edu*	**Does not affect IRB approval.**

PROTOCOL COVER PAGE

Protocol Title	Robotic Inference of Gestural Indication
Principal Investigator/Faculty Advisor	Name: Junaed Sattar
	Department: Computer Science and Engineering
	Telephone Number:
	Email Address:
Student Investigator	Name: Andrea Walker
	Current Academic Status (Student, Fellow, Resident):Student
	Department: Computer Science and Engineering
	Telephone Number:
	Institutional Email Address:
Student Investigator	Name: Luoyao Chen
	Current Academic Status (Student, Fellow, Resident):Student
	Department:
	Telephone Number:
	Institutional Email Address:
Scientific Assessment	Choose an item.
Version Number/Date:	Version: 1.0 Date: (1/14/2021)

REVISION HISTORY

Revision #	Version Date	Summary of Changes	Consent Change?

NOTE: Leave this section blank for the initial submission. The revision history should be documented for modifications to approved studies.

Table of Contents

1.0Objectives 8
2.0Background 8
3.0Study Endpoints/Events/Outcomes 8
4.0Study Intervention(s)/Interaction(s) 8
5.0Procedures Involved 8
6.0Data Banking 9
7.0Sharing of Results with Participants 9
8.0Study Duration 9
9.0Study Population 9
11.0Number of Participants 12
12.0Recruitment Methods 12
13.0Withdrawal of Participants 14
14.0Risks to Participants 14
15.0Incomplete Disclosure or Deception 14
16.0Potential Benefits to Participants 15
17.0Statistical Considerations 15
18.0Health Information and Privacy Compliance 15
19.0Confidentiality 18
20.0Provisions to Monitor the Data to Ensure the Safety of Participants 19
21.0Compensation for Research-Related Injury 20
22.0Consent Process 20
23.0Setting 23
24.0Multi-Site Research 24
25.0Coordinating Center Research 25
26.0Resources Available 26
27.0References 26

ABBREVIATIONS/DEFINITIONS

- HRI - Human Robot Interaction

1. Objectives

1.1. Purpose: The purpose of this research project is to develop an algorithm for aquatic robots to infer the object indicated within a scene and the action that should be taken with the said object when a diver performs a pointing gesture.

2. Background

2.1. Significance of Research Question/Purpose: The problem of creating an algorithm to identify a pointing gesture and interpret the object indicated or intent behind the gesture is a widely investigated topic in terrestrial environments. However, this same problem has not been addressed in the underwater domain. Thus, we intend to bridge this gap in the existing literature and create an algorithm effective in an underwater environment.

2.2. Preliminary Data: Existing publicly available datasets terrestrial pointing gestures, a terrestrial dataset previously created under study IRB (STUDY00011504), and data collected during the PI's underwater (ocean and pool) field trials serve as initial data which may be used during this project.

2.3. Existing Literature: The problem of creating a computer vision algorithm to identify a pointing gesture in a terrestrial environment is a thoroughly investigated topic; and the larger problem of identifying both the pointing gesture and interpreting the object indicated or the intent behind the gesture is addressed in [1][2][3]. This research project contributes to the existing literature by solving this problem in a new domain, underwater. Up until now, no research successfully addressing this topic has been published for the underwater domain.

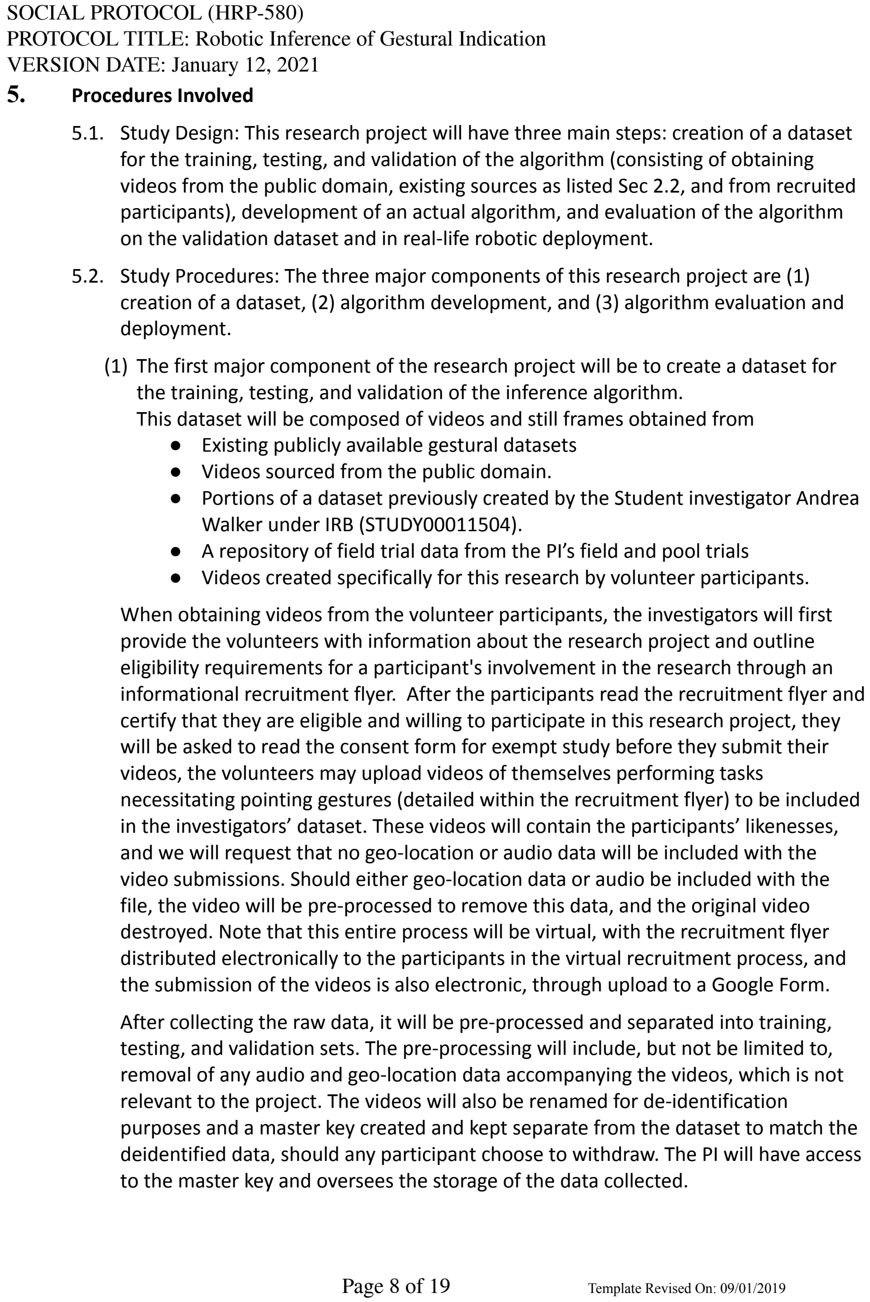
3. Study Endpoints/Events/Outcomes

3.1. Primary Endpoint/Event/Outcome: The primary outcome of this research project will be the development of a computer vision algorithm for aquatic robots which can identify when a diver performs a pointing gesture and subsequently infer the object indicated within a scene along with the action that should be taken with the said object.

3.2. Secondary Endpoint(s)/Event(s)/Outcome(s): Any computer vision algorithm requires a dataset for evaluation and / or training. In order to create the algorithm described above, we will also build a dataset suited to the purpose.

4. Study Intervention(s)/Interaction(s)

4.1. Description: This research project does not include any interactions; however, it includes an intervention in that the participants contributing data will be asked to perform a semi-staged task (such as giving a tour or imitating a cooking show) which will be designed to necessitate the use of pointing gestures. This is the extent of the participants' involvement in this project. The videos submitted by the participants will subsequently be used to develop the investigators' model.



5. Procedures Involved

5.1. Study Design: This research project will have three main steps: creation of a dataset for the training, testing, and validation of the algorithm (consisting of obtaining videos from the public domain, existing sources as listed Sec 2.2, and from recruited participants), development of an actual algorithm, and evaluation of the algorithm on the validation dataset and in real-life robotic deployment.

5.2. Study Procedures: The three major components of this research project are (1) creation of a dataset, (2) algorithm development, and (3) algorithm evaluation and deployment.

(1) The first major component of the research project will be to create a dataset for the training, testing, and validation of the inference algorithm.
This dataset will be composed of videos and still frames obtained from

- Existing publicly available gestural datasets
- Videos sourced from the public domain.
- Portions of a dataset previously created by the Student investigator Andrea Walker under IRB (STUDY00011504).
- A repository of field trial data from the PI's field and pool trials
- Videos created specifically for this research by volunteer participants.

When obtaining videos from the volunteer participants, the investigators will first provide the volunteers with information about the research project and outline eligibility requirements for a participant's involvement in the research through an informational recruitment flyer. After the participants read the recruitment flyer and certify that they are eligible and willing to participate in this research project, they will be asked to read the consent form for exempt study before they submit their videos, the volunteers may upload videos of themselves performing tasks necessitating pointing gestures (detailed within the recruitment flyer) to be included in the investigators' dataset. These videos will contain the participants' likenesses, and we will request that no geo-location or audio data will be included with the video submissions. Should either geo-location data or audio be included with the file, the video will be pre-processed to remove this data, and the original video destroyed. Note that this entire process will be virtual, with the recruitment flyer distributed electronically to the participants in the virtual recruitment process, and the submission of the videos is also electronic, through upload to a Google Form.

After collecting the raw data, it will be pre-processed and separated into training, testing, and validation sets. The pre-processing will include, but not be limited to, removal of any audio and geo-location data accompanying the videos, which is not relevant to the project. The videos will also be renamed for de-identification purposes and a master key created and kept separate from the dataset to match the deidentified data, should any participant choose to withdraw. The PI will have access to the master key and oversees the storage of the data collected.

(2) The second major component of the research project will be development of the computer vision algorithm to identify a pointing gesture and infer both the object indicated within a scene and the action that should be taken with the said object. The development of this algorithm will be using either traditional Computer Vision or Machine / Deep Learning techniques. Depending on the approach utilized, the algorithm may be trained (algorithm parameters set) using the dataset from step (1) and/or experimental fine-tuning.

(3) The third major component of the research project will be evaluation and deployment of the algorithm developed in step (2). This evaluation will be made using the dataset from (1), using metrics suitable for the algorithm designed. Deployment of the algorithm will be made on actual robots within the IRV research lab. This third portion component of the project does not specifically study the human interaction aspect of this project (addressed in step (1) and (2); rather it is an evaluation of the efficacy of the model).

5.3. Follow-Up: N/A There will be no follow-up data collected from volunteer participants in this research project.

5.4. Individually Identifiable Health Information: N/A There will be no medical information collected.

6. **Data Banking**

6.1. Storage and Access: At the conclusion of this specific research project, the dataset created as described in step (1) within section 5.2 will be stored for future use by the IRVLab. This dataset will be stored in Box Secure storage, which requires two-factor authentication to access, and access will only be granted to researchers within the IRVLab whose research depends on it, as determined by the PI.

6.2. Data: The entire dataset developed in step (1) within section 5.2 will be banked for future use. Specifically, the final processed dataset, separated into training, testing, and validation sets will be stored.
Further, the data used to create this dataset, including

- Raw frames or videos from existing publicly available gestural datasets
- Raw Videos sourced from the public domain.
- The dataset previously created by the Student investigator Andrea Walker under IRB (STUDY00011504).
- Raw Videos and frames from a repository of field trial data from the PI's field and pool trials
- Pre-processed videos contributed to this research project by volunteer participants. **Note:** the videos contributed to this research project will be banked in a post-processed form where any audio and/or geolocation data has been removed. Otherwise, these banked videos will be the original participants' submissions and will be banked for future use.

6.3. Release/Sharing: The sharing of this data will be restricted to IRVLab members who have completed all necessary training and their research requires such data, as determined by the PI.

7. Sharing of Results with Participants

7.1. N/A Data will not be shared with the participants.

8. Study Duration

- Participants' choice, anticipated duration minimum of 10 minutes, maximum of 1 hour.
- Participants may be recruited for up to 6 months.
- This research project is anticipated to be complete by June 2021.

9. Study Population

9.1. Inclusion Criteria: Any participant over the age of 18 who is both willing and physically able to record a video of themselves performing pointing gestures will be eligible to participate. This permit but does not target the inclusion of participants who are pregnant and above the age of 18. Further, since some students over the age of 18 may be standing members of the National Guard, we do include Active members of the military (service members) and DoD personnel (including civilian employees) in the included population. Since these two groups of participants are not targeted and their participation in this minimal risk research project does not affect their status as part of these vulnerable populations, there should be no reason to restrict their participation. Lastly, since the students and employees of the investigators have an indirect personal interest and can make uniquely valuable contributions to this research project, they are included in this research project, with the condition that their terms of employment or student status will not be affected in any way due to their participation or non-participation in the research project.

9.2. Exclusion Criteria: Any person under the age of 18, or who is either unwilling or physically unable to record a video of themselves performing pointing gestures will be ineligible to participate. In addition, any vulnerable population not mentioned in the inclusion criteria above will be excluded.

9.3. Screening: Potential participants will be screened first when they contact the researcher listed on the flyer for eligibility and then again through consenting in which they must certify that they are above the age of 18 and willing for the investigators to use their video submissions for data.

10. Vulnerable Populations

10.1. Vulnerable Populations:

Population / Group	Identify whether any of the following populations will be targeted, included (not necessarily targeted) or excluded from participation in the study.
Children	Not included
Pregnant women/fetuses/neonates	Included
Prisoners	Not included
Adults lacking capacity to consent and/or adults with diminished capacity to consent, including, but not limited to, those with acute medical conditions, psychiatric disorders, neurologic disorders, developmental disorders, and behavioral disorders	Not included
Non-English speakers	Not included
Those unable to read (illiterate)	Not included
Employees of the researcher	Included
Students of the researcher	Included
Undervalued or disenfranchised social group	Not included
Active members of the military (service members), DoD personnel (including civilian employees)	Included
Individual or group that is approached for participation in research during a stressful situation such as emergency room setting, childbirth (labor), etc.	Not included
Individual or group that is disadvantaged in the distribution of	Not included

social goods and services such as income, housing, or healthcare.	
Individual or group with a serious health condition for which there are no satisfactory standard treatments.	Not included
Individual or group with a fear of negative consequences for not participating in the research (e.g. institutionalization, deportation, disclosure of stigmatizing behavior).	Not included
Any other circumstance/dynamic that could increase vulnerability to coercion or exploitation that might influence consent to research or decision to continue in research.	Not included

1.1. 10.2 Additional Safeguards: This permit but does not target the inclusion of participants who are pregnant and above the age of 18. Further, since some students over the age of 18 may be standing members of the National Guard, we do include Active members of the military (service members) and DoD personnel (including civilian employees) in the included population. Since these two groups of participants are not targeted and their participation in this minimal risk research project does not affect their status as part of these vulnerable populations, there should be no reason to restrict their participation. Lastly, since the students and employees of the investigators have an indirect personal interest and can make uniquely valuable contributions to this research project, they are included in this research project, with the condition that their terms of employment or student status will not be affected in any way due to their participation or non-participation in the research project.

2. Number of Participants

2.1. Number of Participants to be Consented: Approximately 75 participants will be recruited to contribute to the dataset for this research project.

3. Recruitment Methods

3.1. Recruitment Process: The participants will be recruited from the time of IRB approval through June 2021. They will be recruited on virtual platforms including through email, social media platforms, graduate student associations and through word of mouth of the acquaintances of the researchers.

3.2. Source of Participants: The investigators' department, fellow students in the graduate student association, and associates of the researchers.

3.3. Identification of Potential Participants: The participants will self-identify after reading the recruitment materials created and distributed as described in sections 12.1 and 12.2 .

3.4. Recruitment Materials: Web-posting and Flyers will be the main recruitment materials.

3.5. Payment: Each participant contributing a video to the research project will receive a $10 Amazon e-gift card as a thank you gift. Each participant will receive a maximum of one (1) $10 Amazon gift card, regardless of the number of video submissions contributed to the research project. Receiving the thank you gift card is contingent upon the video submission meeting the study criteria, and the gift card will be electronically delivered to the eligible participants via the email they provide with their video submission on the intake form.

4. Withdrawal of Participants

4.1. Withdrawal Circumstances: If the videos provided by the participants do not contain any useful information (i.e., do not conform to the video requirements listed in the recruitment materials), they will be removed from the dataset. Participants will be notified via the email they provide on the intake form if their video is withdrawn from the research project, and will have the option of submitting another video for consideration if they wish.

4.2. Withdrawal Procedures: If a participant wishes to withdraw from the research project, they must submit their withdrawal in writing, their video submissions will be deleted permanently from the dataset. However, if a model has already been trained at the time of the withdrawal, the model trained using their data will be retained, since the model itself is generalized and data on the withdrawn participant cannot be directly extracted from the model.

4.3. Termination Procedures: PI will have the discretion to terminate any participants who do not comply with the study requirements.

5. Risks to Participants

5.1. Foreseeable Risks: Participation in this project is voluntary and as such bears minimal risk to the participant. The dataset collected only contains the participants' likenesses, but the dataset will be deidentified so that the participants' names are not revealed by any naming scheme. A master key will be kept separate from the dataset which matches the deidentified data with the corresponding consent form, should any participant choose to withdraw. There is no further personal information about the participants stored, so there is no social, legal, psychological or economic risk. Any risk of physical injury is not introduced by the study directly and would only be introduced by the actions or choices of the participants themselves.

5.2. Reproduction Risks: N/A

5.3. Risks to Others: N/A

6. Incomplete Disclosure or Deception

6.1. Incomplete Disclosure or Deception: N/A

7. Potential Benefits to Participants

7.1. Potential Benefits: There are no guaranteed benefits to the participants from taking part in this research. We likewise cannot promise any benefits to other parties from your taking part in this research.

8. Statistical Considerations

8.1. Data Analysis Plan: The data analysis process will begin with pre-processing to remove audio and geo-location data, if included with the videos contributed to the dataset. After pre-processing, the video data will be analyzed through manual annotation of the video submissions, labeling each frame. The videos may also be resized to a standard shape and normalized. After this dataset preparation, the data will be separated into training, test, and validation sets.

8.2. Power Analysis: N/A

8.3. Statistical Analysis: N/A

8.4. Data Integrity: The quality of the data will be determined manually, and any blurry frames will be removed from the dataset. Similarly, any data that does not conform to the project guidelines (as set out in the recruitment materials) will be removed. The PI will maintain the dataset throughout the study.

9. Health Information and Privacy Compliance

N/A

9.1. Select which of the following is applicable to your research:
N/A

☒ My research does not require access to individual health information and therefore assert HIPAA does not apply.

☐ I am requesting that all research participants sign a HIPCO approved HIPAA

Disclosure Authorization to participate in the research (either the standalone form or the combined consent and HIPAA Authorization).

☐ I am requesting the IRB to approve a Waiver or an alteration of research participant authorization to participate in the research.

Appropriate Use for Research:

☐ An external IRB (e.g. Advarra) is reviewing and we are requesting use of the authorization language embedded in the template consent form in lieu of the U of M stand-alone HIPAA Authorization. Note: External IRB must be serving as the privacy board for this option.

9.2. Identify the source of Private Health Information you will be using for your research (Check all that apply)
N/A

☐ I will use the Informatics Consulting Services (ICS) available through CTSI (also referred to as the University's Information Exchange (IE) or data shelter) to pull records for me

☐ I will collect information directly from research participants.

☐ I will use University services to access and retrieve records from the Bone Marrow Transplant (BMPT) database, also known as the HSCT (Hematopoietic Stem Cell Transplant) database.

☐ I will pull records directly from EPIC.

☐ I will retrieve record directly from axiUm / MiPACS

☐ I will receive data from the Center for Medicare/Medicaid Services

☐ I will receive a limited data set from another institution

☐ Other. Describe:

Explain how you will ensure that only records of patients who have agreed to have their information used for research will be reviewed.
N/A

9.3. Approximate number of records required for review:
N/A

9.4. Please describe how you will communicate with research participants during the course of this research. Check all applicable boxes

☐ This research involves record review only. There will be no communication with research participants.

☐ Communication with research participants will take place in the course of treatment, through MyChart, or other similar forms of communication used with patients receiving treatment.

☐ Communication with research participants will take place outside of treatment settings. If this box is selected, please describe the type of communication and how it will be received by participants.

Access to participants
N/A

9.5. Location(s) of storage, sharing and analysis of research data, including any links to research data (check all that apply).
N/A

☐ In the data shelter of the Information Exchange (IE)

☐ Store☐ Analyze☐ Share

☐ In the Bone Marrow Transplant (BMT) database, also known as the HSCT (Hematopoietic Stem Cell Transplant) Database

☐ Store☐ Analyze☐ Share

☐ In REDCap (recap.ahc.umn.edu)

☐ Store☐ Analyze☐ Share

☐ In Qualtrics (qualtrics.umn.edu)

☐ Store☐ Analyze☐ Share

☐ In OnCore (oncore.umn.edu)

☐ Store☐ Analyze☐ Share

☐ In the University's Box Secure Storage (box.umn.edu)

☐ Store☐ Analyze☐ Share

☐ In an AHC-IS supported server. Provide folder path, location of server and IT Support Contact:

☐ Store☐ Analyze☐ Share

☐ In an AHC-IS supported desktop or laptop.

Provide UMN device numbers of all devices:

☐ Store☐ Analyze☐ Share

☐ Other.

Indicate if data will be collected, downloaded, accessed, shared or stored using a server, desktop, laptop, external drive or mobile device (including a tablet computer such as an iPad or a smartform (iPhone or Android devices) that you have not already identified in the preceding questions

☐I will use a server not previously listed to collect/download research data

☐I will use a desktop or laptop not previously listed

☐I will use an external hard drive or USB drive ("flash" or "thumb" drives) not previously listed

☐I will use a mobile device such as an tablet or smartphone not previously listed

9.6. Consultants. Vendors. Third Parties. N/A

9.7. Links to identifiable data: N/A

9.8. Sharing of Data with Research Team Members. Team members will be given access to the required data base on their research requirements and PI discretion.

9.9. Storage and Disposal of Paper Documents: Consent forms and videos will be collected electronically and stored.

10. Confidentiality

10.1. Data Security: The PI will oversee the data collected. The data will be stored under UMN approved such as BOX, Redcap or google drive. Only approved research staff will be given access to the data once they have completed all necessary training.

11. Provisions to Monitor the Data to Ensure the Safety of Participants

11.1. Data Integrity Monitoring.

- This is a minimal risk study, and since the dataset will not contain any personally identifiable health information which could pose a risk to the participants, there is no necessity for independent data integrity monitoring.

11.2. Data Safety Monitoring.

- As stated above, this is a minimal risk study, and the data collected will pose no harm or risk to the safety of the participants. Thus, Data safety monitoring is also not applicable.

12. Compensation for Research-Related Injury

12.1. Compensation for Research-Related Injury: There will be no compensation for research-related injury as this is a minimal risk study.

12.2. Contract Language: N/A

13. Consent Process

- When participants volunteer for the research project, they will be asked to read an exempt consent form prior to contributing video data to the project. Participants will not be permitted to proceed with the project unless their consent has been given through clicking to confirm that they are over 18 of age and read the exempt consent form. The entire consent process will take place virtually / in an online environment.
- The waiting time period is entirely dependent on the time it takes the participant to confirm they read and understand the exempt consent form..
- The investigators will always be reachable to the participants over email if they ever want to withdraw their video from the dataset.

13.2. Waiver or Alteration of Consent Process (when consent will not be obtained, required information will not be disclosed, or the research involves deception): It's an exempt study so we will be asking participants to read the exempt study consent form.

13.3. Waiver of Written/Signed Documentation of Consent (when written/signed consent will not be obtained): It's an exempt study so we will be asking participants to read the exempt study consent form.

13.4. Non-English-Speaking Participants: N/A

13.5. Participants Who Are Not Yet Adults (infants, children, teenagers under 18 years of age): N/A

13.6. Cognitively Impaired Adults, or adults with fluctuating or diminished capacity to consent: N/A

13.7. Adults Unable to Consent: N/A

14. Setting

14.1. Research Sites: Online

14.2. International Research: NA

14.3. Community Based Participatory Research: NA

15. **Multi-Site Research:** N/A

16. **Coordinating Center Research:** N/A

17. Resources Available

17.1. Resources Available:

- The PI Junaed Sattar will provide oversight for the duration of the entire research project, as described in section 5.2.
- We have access to approximately 200 suitable participants within the University of Minnesota Computer Science graduate student body. If 2.5% of these potential participants respond to the recruitment, we will have our minimum of 5 recruited participants.
- This research will be conducted from the present until June 2021.
- The facilities used for this research will be primarily virtual, consisting of the internet, the researchers', university-hosted, and IRVLab compute servers. In the final stage of robotic deployment the facilities of the IRVLab may be utilized, in accordance with the guidelines outlined by the Sunrise Plan.
- All potential collaborators assisting with this research will be members of the IRVLab, and if they are granted access to the data collected from participants prior to de-identification, will be required to read this protocol prior to utilizing this data.

18. References

[1] C. P. Quintero, R. Tatsambon, M. Gridseth and M. Jägersand, "Visual pointing gestures for bi-directional human robot interaction in a pick-and-place task," 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, 2015, pp. 349-354, doi: 10.1109/ROMAN.2015.7333604.

[2] D. Shukla, O. Erkent and J. Piater, "Probabilistic Detection of Pointing Directions for Human-Robot Interaction," 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Adelaide, SA, 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371296.

[3] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang and J. Sun, "Learning Human-Object Interaction Detection Using Interaction Points," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 4115-4124, doi: 10.1109/CVPR42600.2020.00417.

A.2 Social-Behavioral Consent Form, HRP-582

Since our study was ultimately determined to be exempt from IRB review, this Consent Form was ultimately replaced with the HRP-587 Information form in §A.4.

Consent Form

Title of Research Study: Robotic Inference of Gestural Indication [protocol #]

Investigator Team Contact Information: **Junaed Sattar**
For questions about research appointments, the research study, research results, or other concerns, call the study team at:

Investigator Name: Junaed Sattar Investigator Departmental Affiliation: Computer Science and Engineering Phone Number: Email Address:	Student Investigator Name: Andrea Walker Phone Number: Email Address: Student Investigator Name: Luoyao Chen Phone Number: Email Address:

Supported By:
This research is supported by IIS #1845364:Towards robust and natural underwater human-robot interaction, CON000000078102.

Key Information About This Research Study

The following is a short summary to help you decide whether or not to be a part of this research study. More detailed information is listed later on in this form. What is research?

- The goal of research is to learn new things in order to help people in the future. Investigators learn things by following the same plan with a number of participants, so they do not usually make changes to the plan for individual research participants. You, as an individual, may or may not be helped by volunteering for a research study.

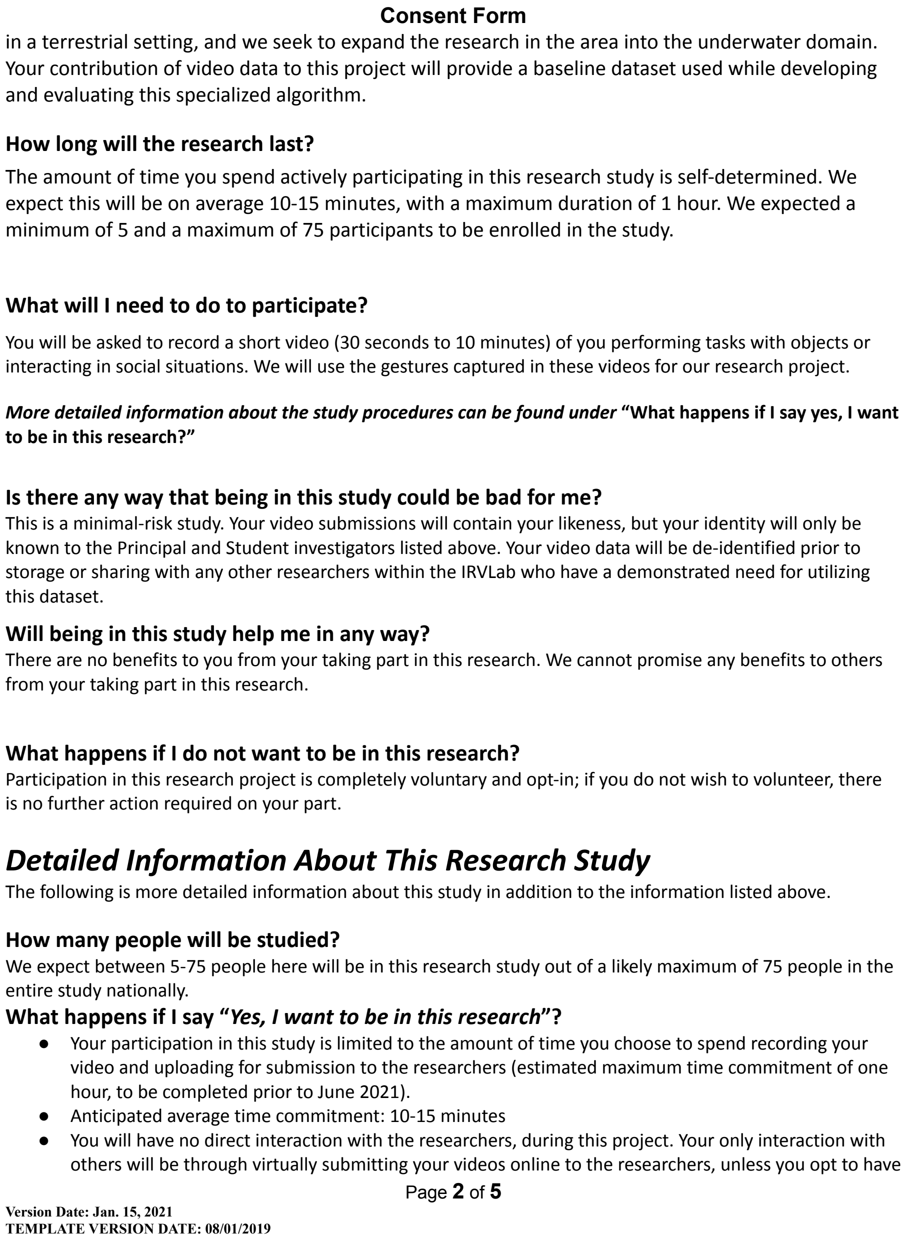
Why am I being invited to take part in this research study?

We are asking you to take part in this research study because the researchers are seeking adults over the age of 18 with full upper-body mobility to create videos of themselves making hand and arm gestures. What should I know about a research study?

- Someone will explain this research study to you.
- Whether or not you take part is up to you.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.

Why is this research being done?

The goal of this research project is to create an algorithm that allows an aquatic robot to identify gestures and infer the intent behind the identified gesture. This problem has been widely investigated



in a terrestrial setting, and we seek to expand the research in the area into the underwater domain. Your contribution of video data to this project will provide a baseline dataset used while developing and evaluating this specialized algorithm.

How long will the research last?

The amount of time you spend actively participating in this research study is self-determined. We expect this will be on average 10-15 minutes, with a maximum duration of 1 hour. We expected a minimum of 5 and a maximum of 75 participants to be enrolled in the study.

What will I need to do to participate?

You will be asked to record a short video (30 seconds to 10 minutes) of you performing tasks with objects or interacting in social situations. We will use the gestures captured in these videos for our research project.

More detailed information about the study procedures can be found under **"What happens if I say yes, I want to be in this research?"**

Is there any way that being in this study could be bad for me?

This is a minimal-risk study. Your video submissions will contain your likeness, but your identity will only be known to the Principal and Student investigators listed above. Your video data will be de-identified prior to storage or sharing with any other researchers within the IRVLab who have a demonstrated need for utilizing this dataset.

Will being in this study help me in any way?

There are no benefits to you from your taking part in this research. We cannot promise any benefits to others from your taking part in this research.

What happens if I do not want to be in this research?

Participation in this research project is completely voluntary and opt-in; if you do not wish to volunteer, there is no further action required on your part.

Detailed Information About This Research Study

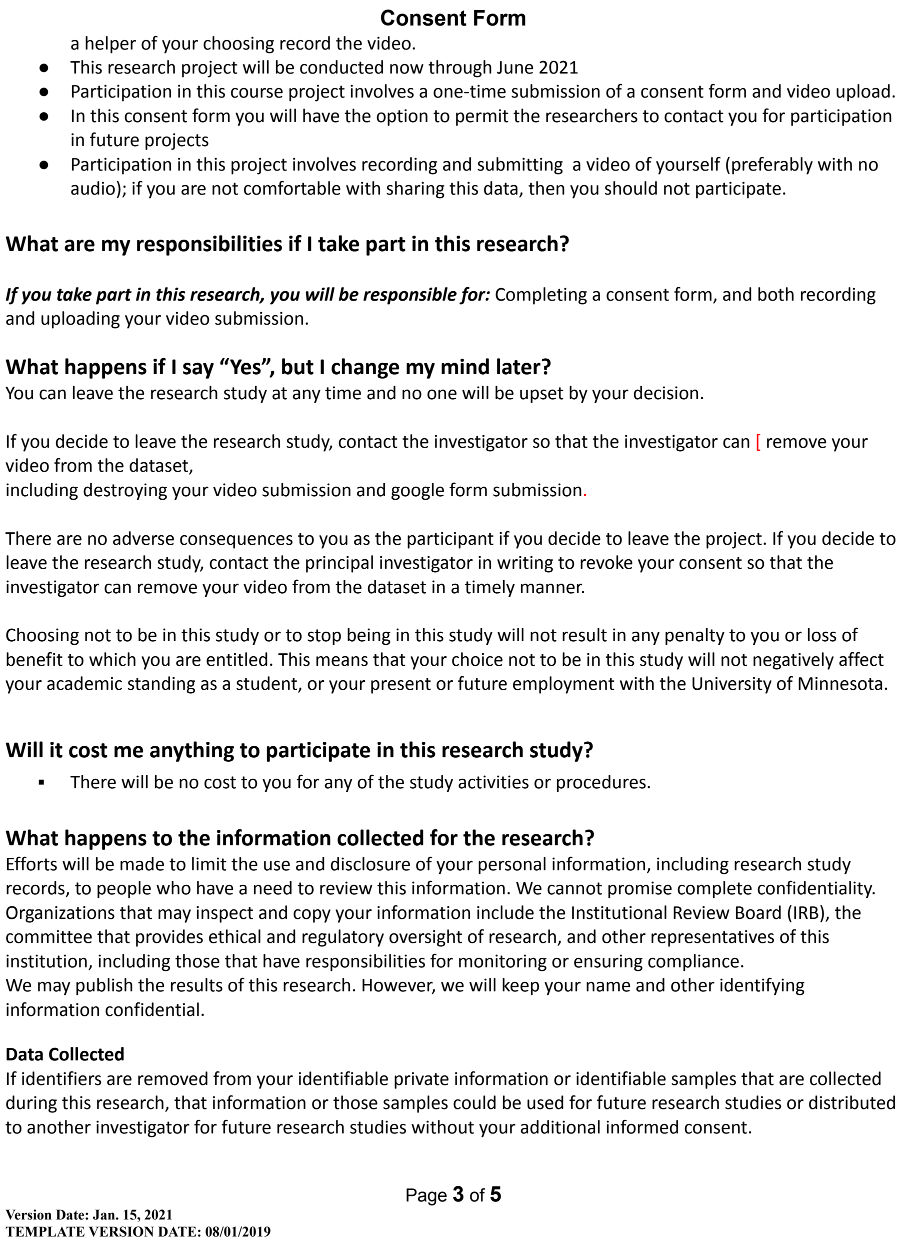
The following is more detailed information about this study in addition to the information listed above.

How many people will be studied?

We expect between 5-75 people here will be in this research study out of a likely maximum of 75 people in the entire study nationally.

What happens if I say "*Yes, I want to be in this research*"?

- Your participation in this study is limited to the amount of time you choose to spend recording your video and uploading for submission to the researchers (estimated maximum time commitment of one hour, to be completed prior to June 2021).
- Anticipated average time commitment: 10-15 minutes
- You will have no direct interaction with the researchers, during this project. Your only interaction with others will be through virtually submitting your videos online to the researchers, unless you opt to have



a helper of your choosing record the video.

- This research project will be conducted now through June 2021
- Participation in this course project involves a one-time submission of a consent form and video upload.
- In this consent form you will have the option to permit the researchers to contact you for participation in future projects
- Participation in this project involves recording and submitting a video of yourself (preferably with no audio); if you are not comfortable with sharing this data, then you should not participate.

What are my responsibilities if I take part in this research?

If you take part in this research, you will be responsible for: Completing a consent form, and both recording and uploading your video submission.

What happens if I say "Yes", but I change my mind later?

You can leave the research study at any time and no one will be upset by your decision.

If you decide to leave the research study, contact the investigator so that the investigator can [remove your video from the dataset,
including destroying your video submission and google form submission.

There are no adverse consequences to you as the participant if you decide to leave the project. If you decide to leave the research study, contact the principal investigator in writing to revoke your consent so that the investigator can remove your video from the dataset in a timely manner.

Choosing not to be in this study or to stop being in this study will not result in any penalty to you or loss of benefit to which you are entitled. This means that your choice not to be in this study will not negatively affect your academic standing as a student, or your present or future employment with the University of Minnesota.

Will it cost me anything to participate in this research study?

- There will be no cost to you for any of the study activities or procedures.

What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. We cannot promise complete confidentiality. Organizations that may inspect and copy your information include the Institutional Review Board (IRB), the committee that provides ethical and regulatory oversight of research, and other representatives of this institution, including those that have responsibilities for monitoring or ensuring compliance.
We may publish the results of this research. However, we will keep your name and other identifying information confidential.

Data Collected

If identifiers are removed from your identifiable private information or identifiable samples that are collected during this research, that information or those samples could be used for future research studies or distributed to another investigator for future research studies without your additional informed consent.

Consent Form

After the research project is completed, the data will be retained (in an anonymized form) for potential use by the investigators and members of the IRVLab whose research requires such a dataset. The data will be stored on the University of Minnesota's secure Box storage, protected by two-factor authentication and with access restricted to the investigators listed above and IRVLab researchers whose work requires such data. The data you provide will be retained indefinitely for this and subsequent related research projects.

What will be done with my data when this study is over?

Secondary (future) research **without identifiers**:
We will use and may share data for future research. They may be shared with researchers/institutions outside of University of Minnesota.   This could include for profit companies. We will not ask for your consent before using or sharing them. We will remove identifiers from your data, which means that nobody who works with them for future research will know who you are. Therefore, you will not receive any results or financial benefit from future research done on your specimens or data.

Whom do I contact if I have questions, concerns or feedback about my experience?

This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

Will I have a chance to provide feedback after the study is over?

The HRPP may ask you to complete a survey that asks about your experience as a research participant. You do not have to complete the survey if you do not want to. If you do choose to complete the survey, your responses will be anonymous.

If you are not asked to complete a survey, but you would like to share feedback, please contact the study team or the HRPP. See the "Investigator Contact Information" of this form for study team contact information and "Whom do I contact if I have questions, concerns or feedback about my experience?" of this form for HRPP contact information.

Can I be removed from the research?

The person in charge of the research study can remove you from the research study without your approval. Possible reasons for removal include improper or broken video submissions or submission of a consent form without a signature.

We will tell you about any new information that may affect your choice to stay in the research.

Will I be compensated for my participation?

If you agree to take part in this research study, you will receive a $10 Amazon e-gift card as a thank you for your

Consent Form

time and effort. Each participant will receive a maximum of one (1) $10 Amazon e-gift card, regardless of the number of video submissions contributed. Receiving the thank you gift card is contingent upon the video submission meeting the study criteria and the consent form being completed and signed. After confirmation of these requirements, the gift card will be electronically delivered to the eligible participants via the email they provide with their video submission on the intake form.

Optional Elements:

The following research activities are optional, meaning that you do not have to agree to them in order to participate in the research study. Please indicate your willingness to participate in these optional activities by placing your initials next to each activity.

Yes, I agree	**No, I disagree**	
__________	_________	The investigator may audio or video record me to aid with data analysis. The investigator will not share these recordings with anyone outside of the immediate study team.
__________	_________	The investigator may audio or video record me for use in scholarly presentations or publications. My identity may be shared as part of this activity, although the investigator will attempt to limit such identification. I understand the risks associated with such identification.
__________	_________	The investigator may contact me in the future to see whether I am interested in participating in other research studies by Junaed Sattar.

Signature Block for Capable Adult:

Your signature documents your permission to take part in this research. You will be provided a copy of this signed document.

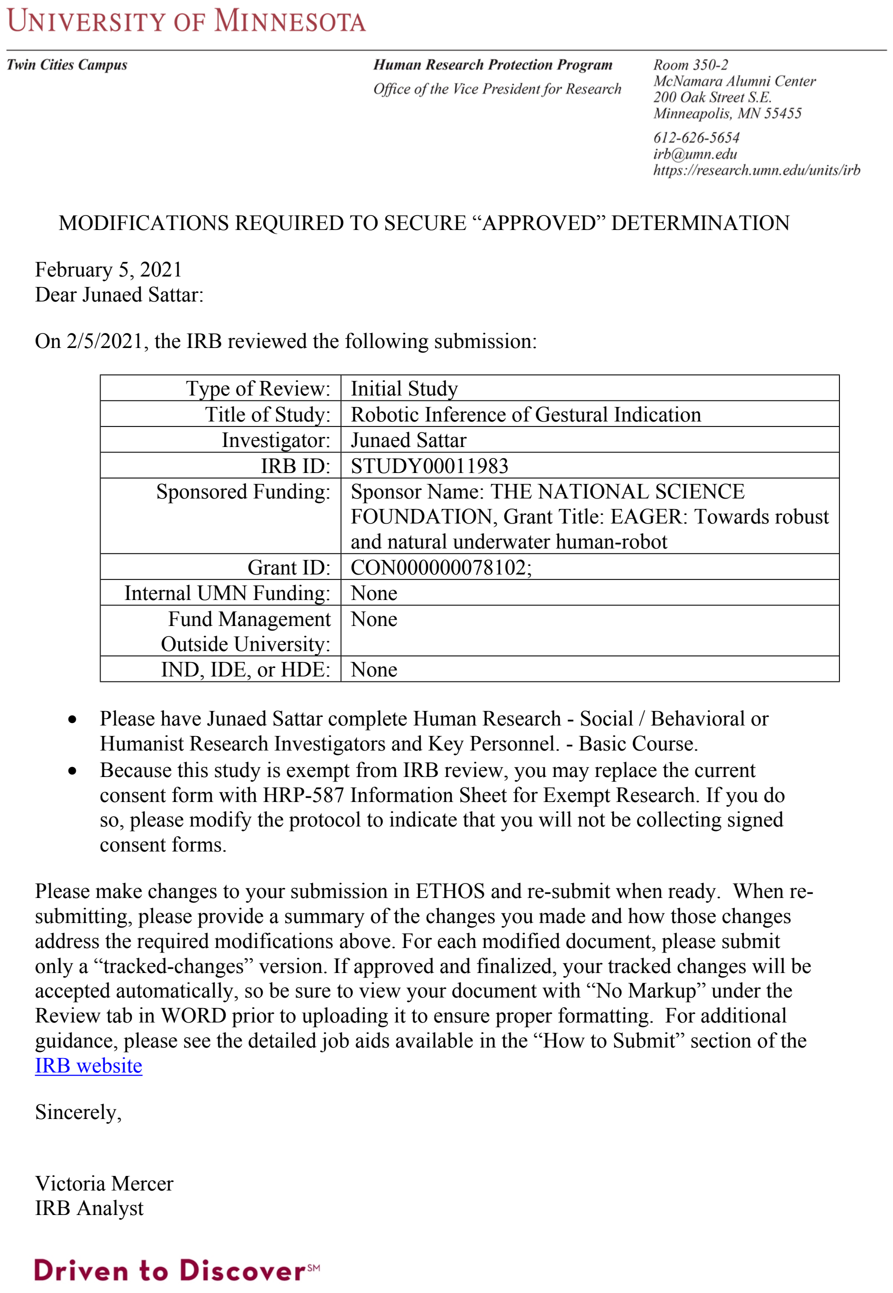
____ ___________________
Signature of Participant Date

____
Printed Name of Participant

____ ___________________
Signature of Person Obtaining Consent Date

____
Printed Name of Person Obtaining Consent

A.3 Exemption Determination



UNIVERSITY OF MINNESOTA

Twin Cities Campus

Human Research Protection Program
Office of the Vice President for Research

Room 350-2
McNamara Alumni Center
200 Oak Street S.E.
Minneapolis, MN 55455

612-626-5654
irb@umn.edu
https://research.umn.edu/units/irb

MODIFICATIONS REQUIRED TO SECURE "APPROVED" DETERMINATION

February 5, 2021
Dear Junaed Sattar:

On 2/5/2021, the IRB reviewed the following submission:

Type of Review:	Initial Study
Title of Study:	Robotic Inference of Gestural Indication
Investigator:	Junaed Sattar
IRB ID:	STUDY00011983
Sponsored Funding:	Sponsor Name: THE NATIONAL SCIENCE FOUNDATION, Grant Title: EAGER: Towards robust and natural underwater human-robot
Grant ID:	CON000000078102;
Internal UMN Funding:	None
Fund Management Outside University:	None
IND, IDE, or HDE:	None

- Please have Junaed Sattar complete Human Research - Social / Behavioral or Humanist Research Investigators and Key Personnel. - Basic Course.
- Because this study is exempt from IRB review, you may replace the current consent form with HRP-587 Information Sheet for Exempt Research. If you do so, please modify the protocol to indicate that you will not be collecting signed consent forms.

Please make changes to your submission in ETHOS and re-submit when ready. When re-submitting, please provide a summary of the changes you made and how those changes address the required modifications above. For each modified document, please submit only a "tracked-changes" version. If approved and finalized, your tracked changes will be accepted automatically, so be sure to view your document with "No Markup" under the Review tab in WORD prior to uploading it to ensure proper formatting. For additional guidance, please see the detailed job aids available in the "How to Submit" section of the IRB website

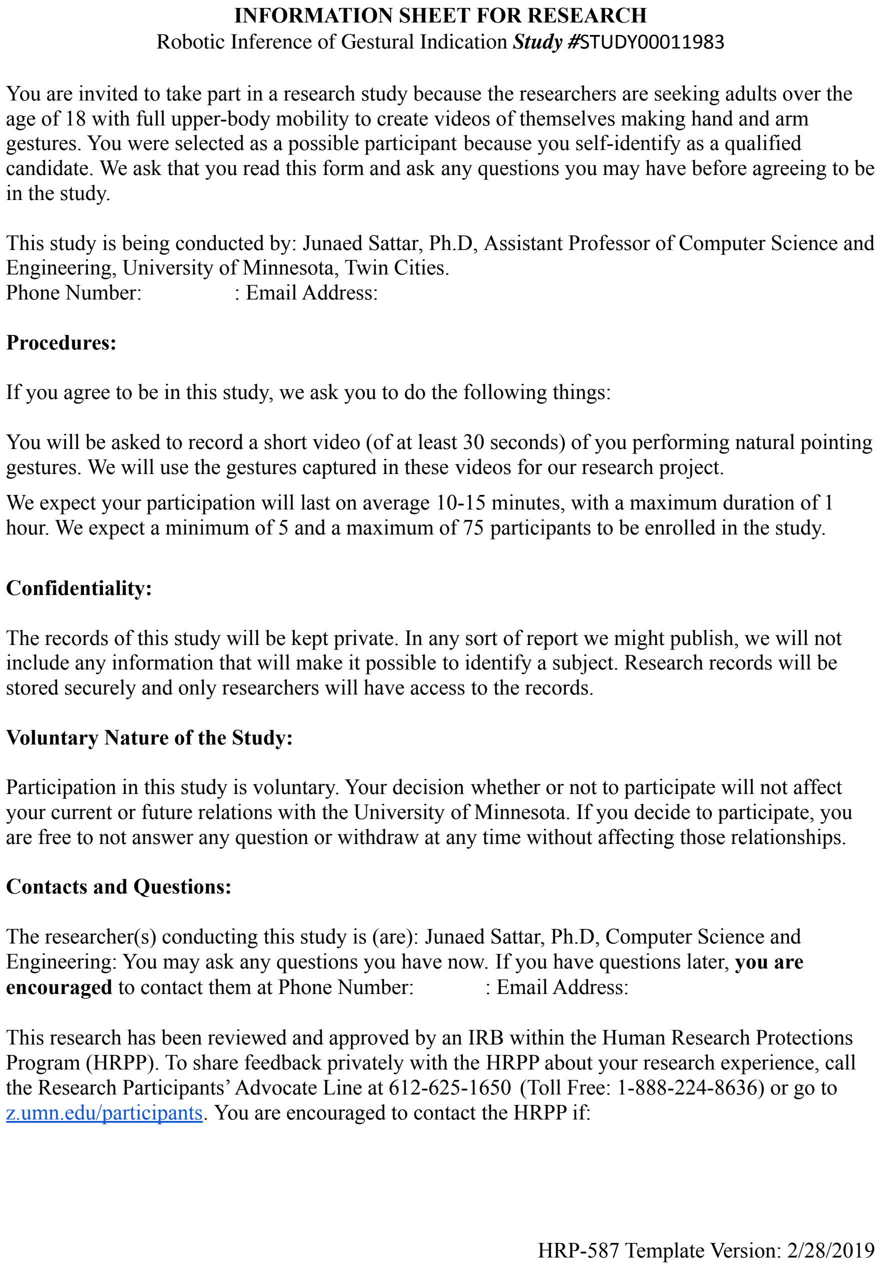
Sincerely,

Victoria Mercer
IRB Analyst

Driven to Discover℠

A.4 Information Sheet for Research, HRP-587

This Information Sheet replaced the Consent Form when our study was determined to be IRB review exempt.



INFORMATION SHEET FOR RESEARCH

Robotic Inference of Gestural Indication ***Study #***STUDY00011983

You are invited to take part in a research study because the researchers are seeking adults over the age of 18 with full upper-body mobility to create videos of themselves making hand and arm gestures. You were selected as a possible participant because you self-identify as a qualified candidate. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Junaed Sattar, Ph.D, Assistant Professor of Computer Science and Engineering, University of Minnesota, Twin Cities.
Phone Number: : Email Address:

Procedures:

If you agree to be in this study, we ask you to do the following things:

You will be asked to record a short video (of at least 30 seconds) of you performing natural pointing gestures. We will use the gestures captured in these videos for our research project.

We expect your participation will last on average 10-15 minutes, with a maximum duration of 1 hour. We expect a minimum of 5 and a maximum of 75 participants to be enrolled in the study.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions:

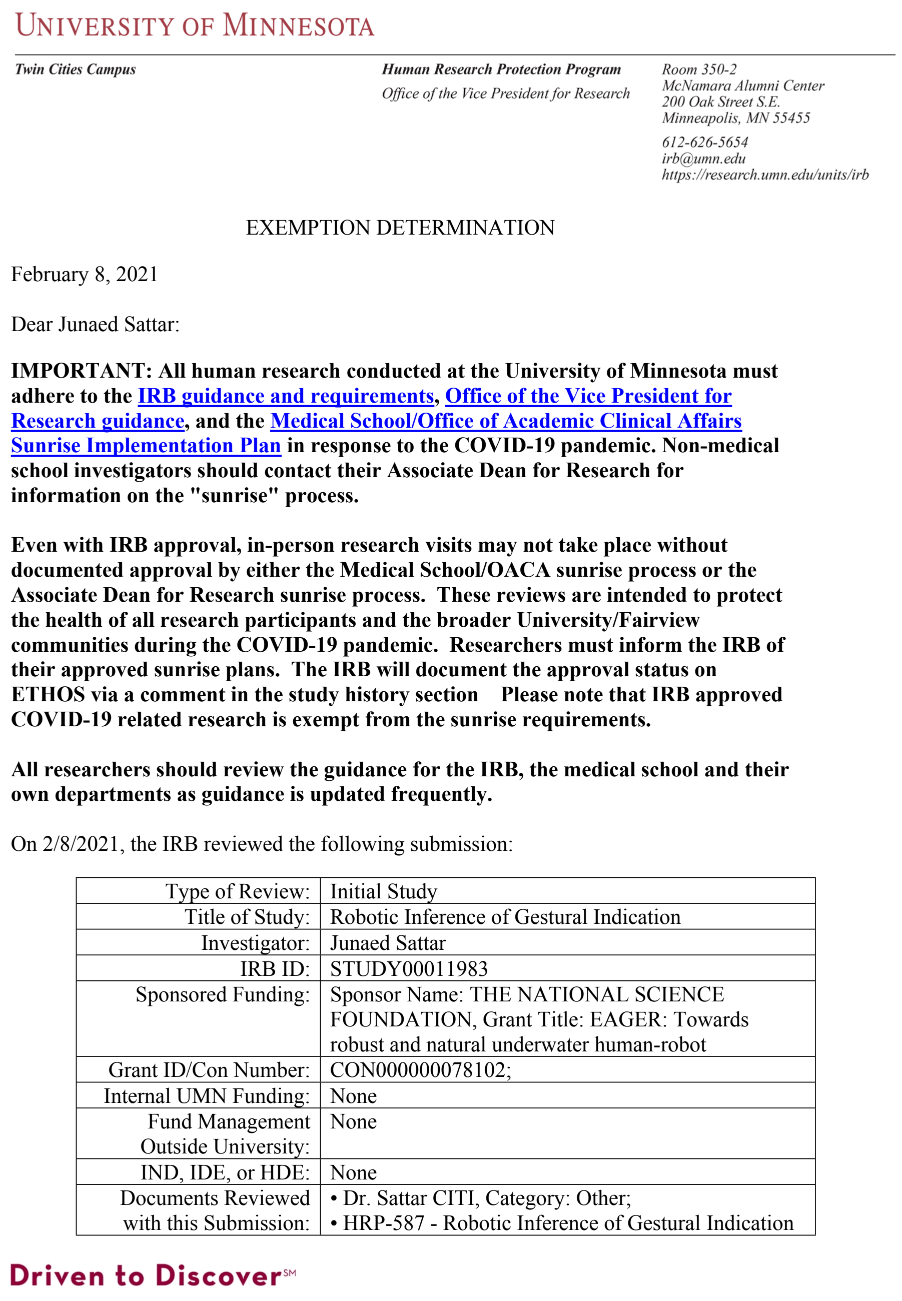
The researcher(s) conducting this study is (are): Junaed Sattar, Ph.D, Computer Science and Engineering: You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at Phone Number: : Email Address:

This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be able to download a copy of this information to keep for your records.

A.5 Final Approval Document



UNIVERSITY OF MINNESOTA

Twin Cities Campus

Human Research Protection Program
Office of the Vice President for Research

Room 350-2
McNamara Alumni Center
200 Oak Street S.E.
Minneapolis, MN 55455

612-626-5654
irb@umn.edu
https://research.umn.edu/units/irb

EXEMPTION DETERMINATION

February 8, 2021

Dear Junaed Sattar:

IMPORTANT: All human research conducted at the University of Minnesota must adhere to the IRB guidance and requirements, Office of the Vice President for Research guidance, and the Medical School/Office of Academic Clinical Affairs Sunrise Implementation Plan in response to the COVID-19 pandemic. Non-medical school investigators should contact their Associate Dean for Research for information on the "sunrise" process.

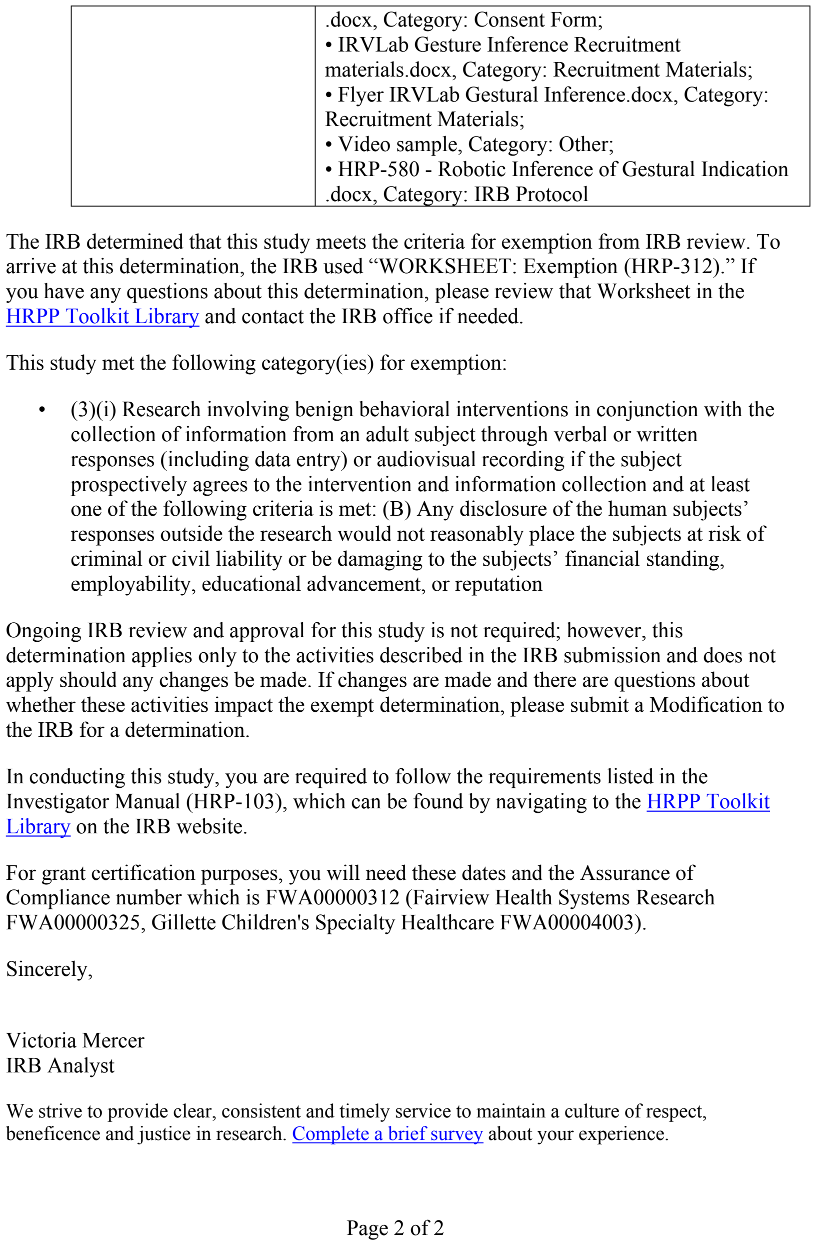
Even with IRB approval, in-person research visits may not take place without documented approval by either the Medical School/OACA sunrise process or the Associate Dean for Research sunrise process. These reviews are intended to protect the health of all research participants and the broader University/Fairview communities during the COVID-19 pandemic. Researchers must inform the IRB of their approved sunrise plans. The IRB will document the approval status on ETHOS via a comment in the study history section Please note that IRB approved COVID-19 related research is exempt from the sunrise requirements.

All researchers should review the guidance for the IRB, the medical school and their own departments as guidance is updated frequently.

On 2/8/2021, the IRB reviewed the following submission:

Type of Review:	Initial Study
Title of Study:	Robotic Inference of Gestural Indication
Investigator:	Junaed Sattar
IRB ID:	STUDY00011983
Sponsored Funding:	Sponsor Name: THE NATIONAL SCIENCE FOUNDATION, Grant Title: EAGER: Towards robust and natural underwater human-robot
Grant ID/Con Number:	CON000000078102;
Internal UMN Funding:	None
Fund Management Outside University:	None
IND, IDE, or HDE:	None
Documents Reviewed with this Submission:	• Dr. Sattar CITI, Category: Other; • HRP-587 - Robotic Inference of Gestural Indication

Driven to Discover℠



	.docx, Category: Consent Form; • IRVLab Gesture Inference Recruitment materials.docx, Category: Recruitment Materials; • Flyer IRVLab Gestural Inference.docx, Category: Recruitment Materials; • Video sample, Category: Other; • HRP-580 - Robotic Inference of Gestural Indication .docx, Category: IRB Protocol

The IRB determined that this study meets the criteria for exemption from IRB review. To arrive at this determination, the IRB used “WORKSHEET: Exemption (HRP-312).” If you have any questions about this determination, please review that Worksheet in the HRPP Toolkit Library and contact the IRB office if needed.

This study met the following category(ies) for exemption:

- (3)(i) Research involving benign behavioral interventions in conjunction with the collection of information from an adult subject through verbal or written responses (including data entry) or audiovisual recording if the subject prospectively agrees to the intervention and information collection and at least one of the following criteria is met: (B) Any disclosure of the human subjects’ responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects’ financial standing, employability, educational advancement, or reputation

Ongoing IRB review and approval for this study is not required; however, this determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit a Modification to the IRB for a determination.

In conducting this study, you are required to follow the requirements listed in the Investigator Manual (HRP-103), which can be found by navigating to the HRPP Toolkit Library on the IRB website.

For grant certification purposes, you will need these dates and the Assurance of Compliance number which is FWA00000312 (Fairview Health Systems Research FWA00000325, Gillette Children's Specialty Healthcare FWA00004003).

Sincerely,

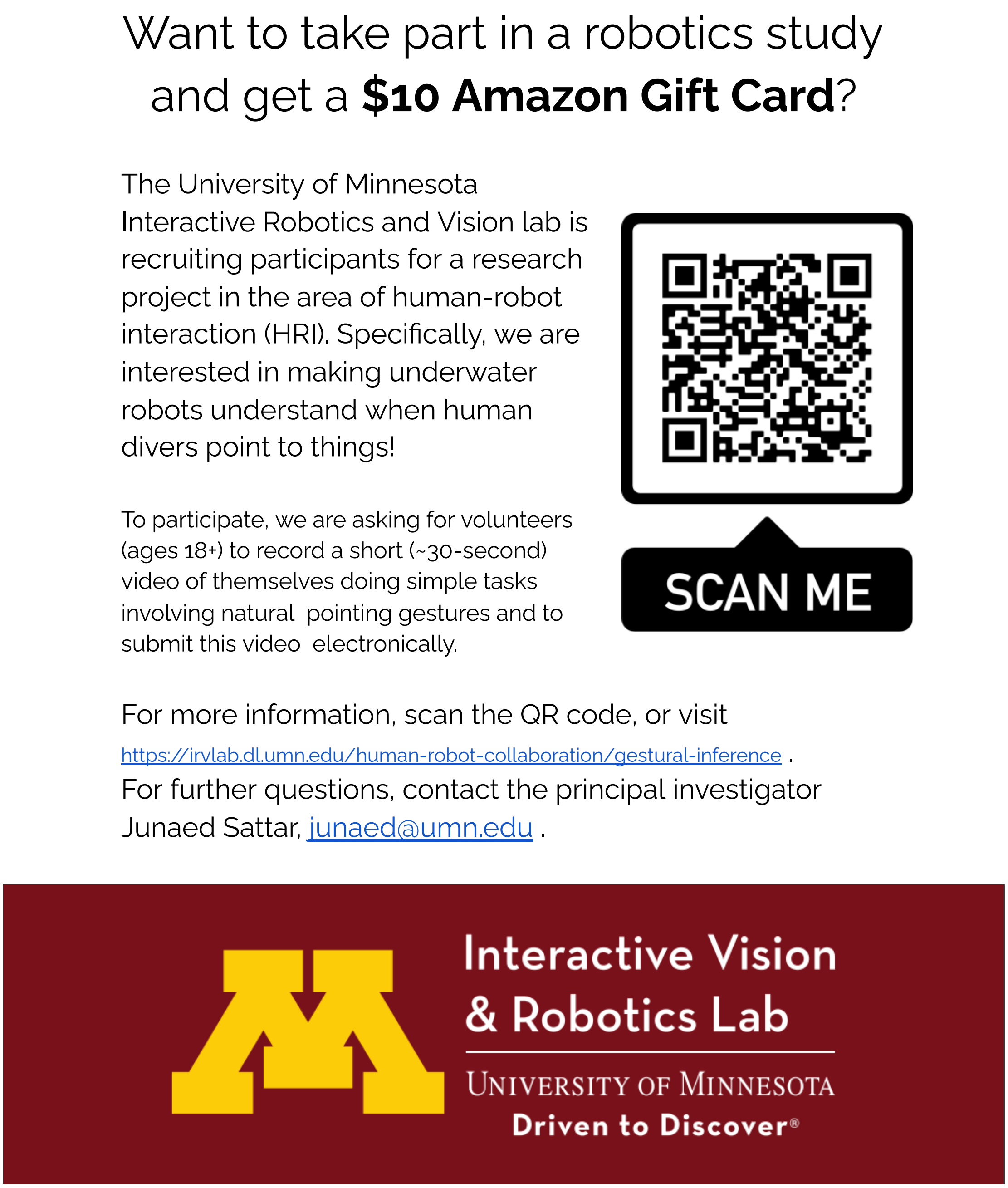
Victoria Mercer
IRB Analyst

We strive to provide clear, consistent and timely service to maintain a culture of respect, beneficence and justice in research. Complete a brief survey about your experience.

Page 2 of 2

A.6 Promotional Materials

Flyer



Want to take part in a robotics study and get a **$10 Amazon Gift Card**?

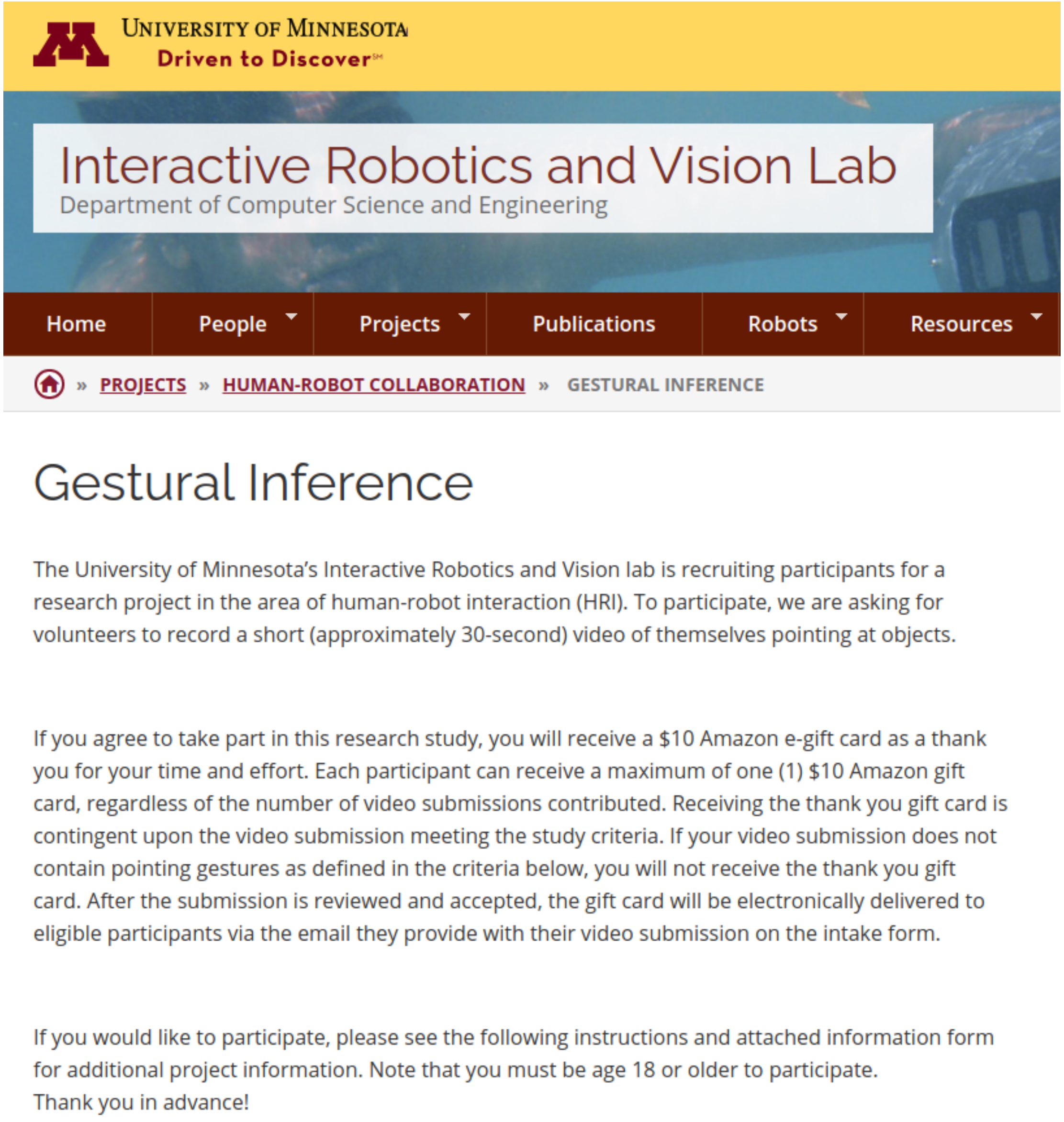
The University of Minnesota Interactive Robotics and Vision lab is recruiting participants for a research project in the area of human-robot interaction (HRI). Specifically, we are interested in making underwater robots understand when human divers point to things!

To participate, we are asking for volunteers (ages 18+) to record a short (~30-second) video of themselves doing simple tasks involving natural pointing gestures and to submit this video electronically.

For more information, scan the QR code, or visit https://irvlab.dl.umn.edu/human-robot-collaboration/gestural-inference .

For further questions, contact the principal investigator Junaed Sattar, junaed@umn.edu .

Website Landing Page


University of Minnesota
Driven to Discover℠
Interactive Robotics and Vision Lab
Department of Computer Science and Engineering
Home
People
Projects
Publications
Robots
Resources
» PROJECTS » HUMAN-ROBOT COLLABORATION » GESTURAL INFERENCE

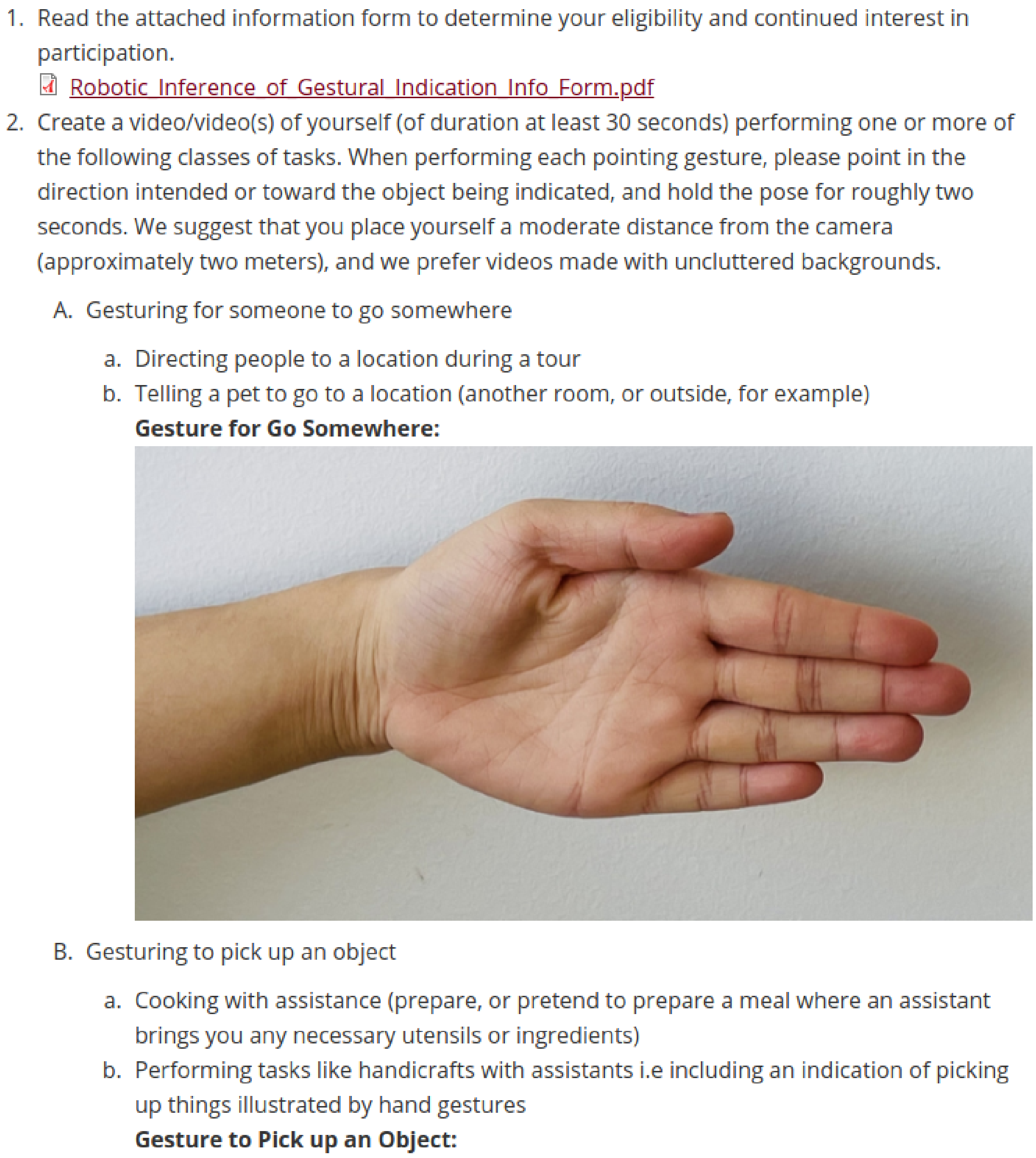
Gestural Inference

The University of Minnesota's Interactive Robotics and Vision lab is recruiting participants for a research project in the area of human-robot interaction (HRI). To participate, we are asking for volunteers to record a short (approximately 30-second) video of themselves pointing at objects.

If you agree to take part in this research study, you will receive a $10 Amazon e-gift card as a thank you for your time and effort. Each participant can receive a maximum of one (1) $10 Amazon gift card, regardless of the number of video submissions contributed. Receiving the thank you gift card is contingent upon the video submission meeting the study criteria. If your video submission does not contain pointing gestures as defined in the criteria below, you will not receive the thank you gift card. After the submission is reviewed and accepted, the gift card will be electronically delivered to eligible participants via the email they provide with their video submission on the intake form.

If you would like to participate, please see the following instructions and attached information form for additional project information. Note that you must be age 18 or older to participate.
Thank you in advance!

Figure A.1: Website part 1.



1. Read the attached information form to determine your eligibility and continued interest in participation.
   Robotic_Inference_of_Gestural_Indication_Info_Form.pdf
2. Create a video/video(s) of yourself (of duration at least 30 seconds) performing one or more of the following classes of tasks. When performing each pointing gesture, please point in the direction intended or toward the object being indicated, and hold the pose for roughly two seconds. We suggest that you place yourself a moderate distance from the camera (approximately two meters), and we prefer videos made with uncluttered backgrounds.

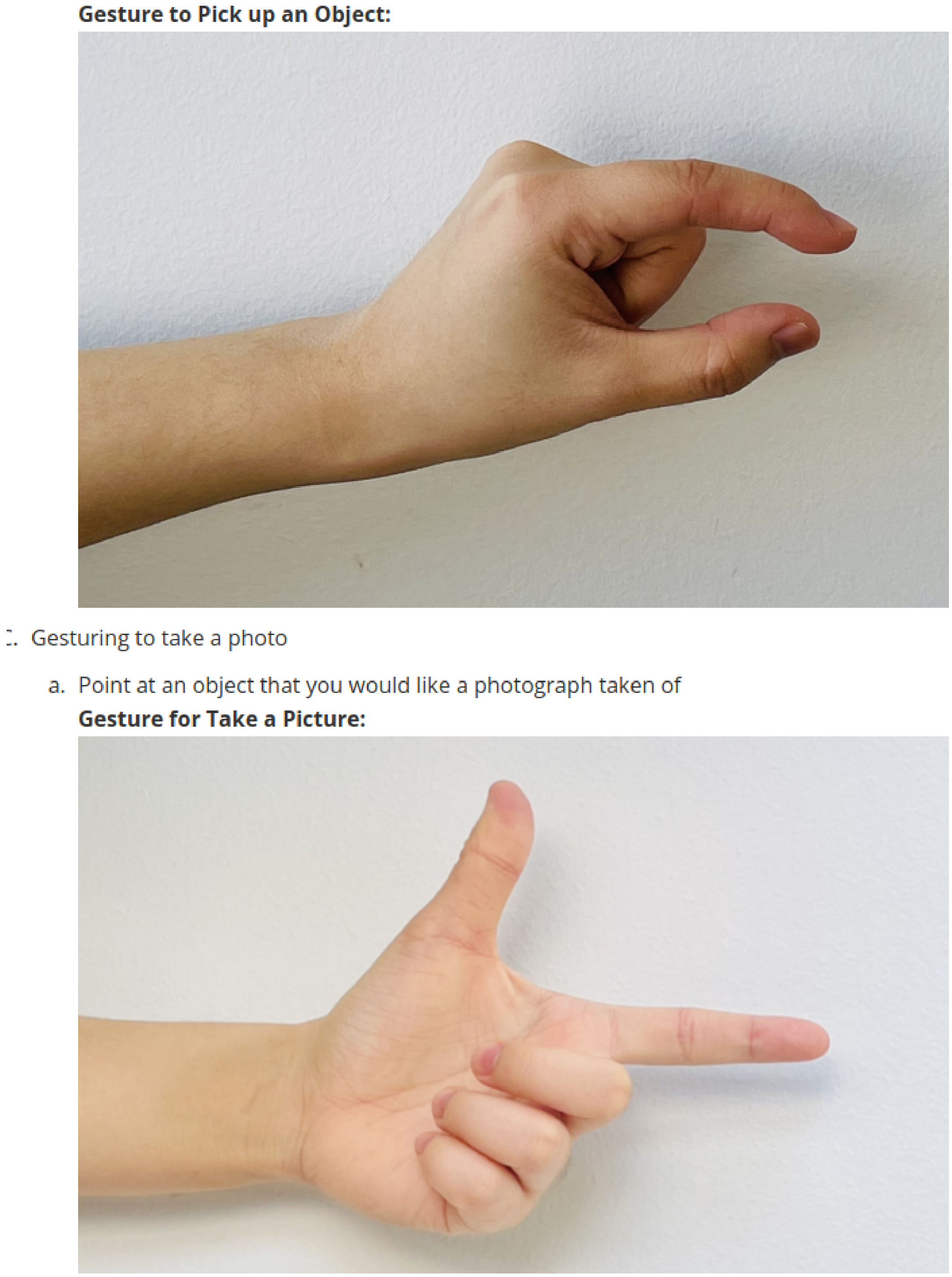
   A. Gesturing for someone to go somewhere

      a. Directing people to a location during a tour
      b. Telling a pet to go to a location (another room, or outside, for example)
      Gesture for Go Somewhere:

   B. Gesturing to pick up an object

      a. Cooking with assistance (prepare, or pretend to prepare a meal where an assistant brings you any necessary utensils or ingredients)
      b. Performing tasks like handicrafts with assistants i.e including an indication of picking up things illustrated by hand gestures
      Gesture to Pick up an Object:

Figure A.2: Website part 2.



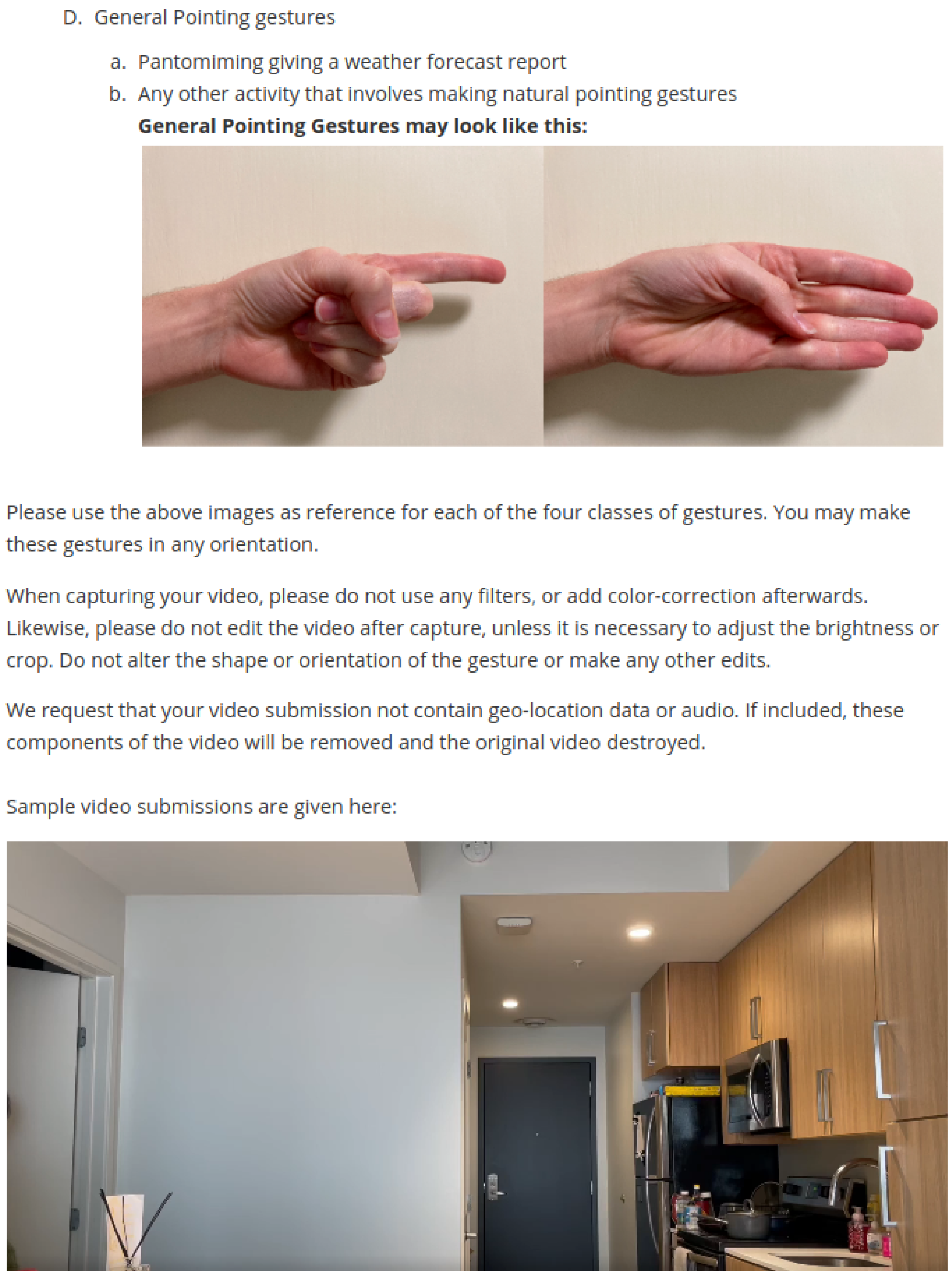
Gesture to Pick up an Object:

Gesturing to take a photo

a. Point at an object that you would like a photograph taken of

Gesture for Take a Picture:

Figure A.3: Website part 3.



D. General Pointing gestures

a. Pantomiming giving a weather forecast report
b. Any other activity that involves making natural pointing gestures

General Pointing Gestures may look like this:

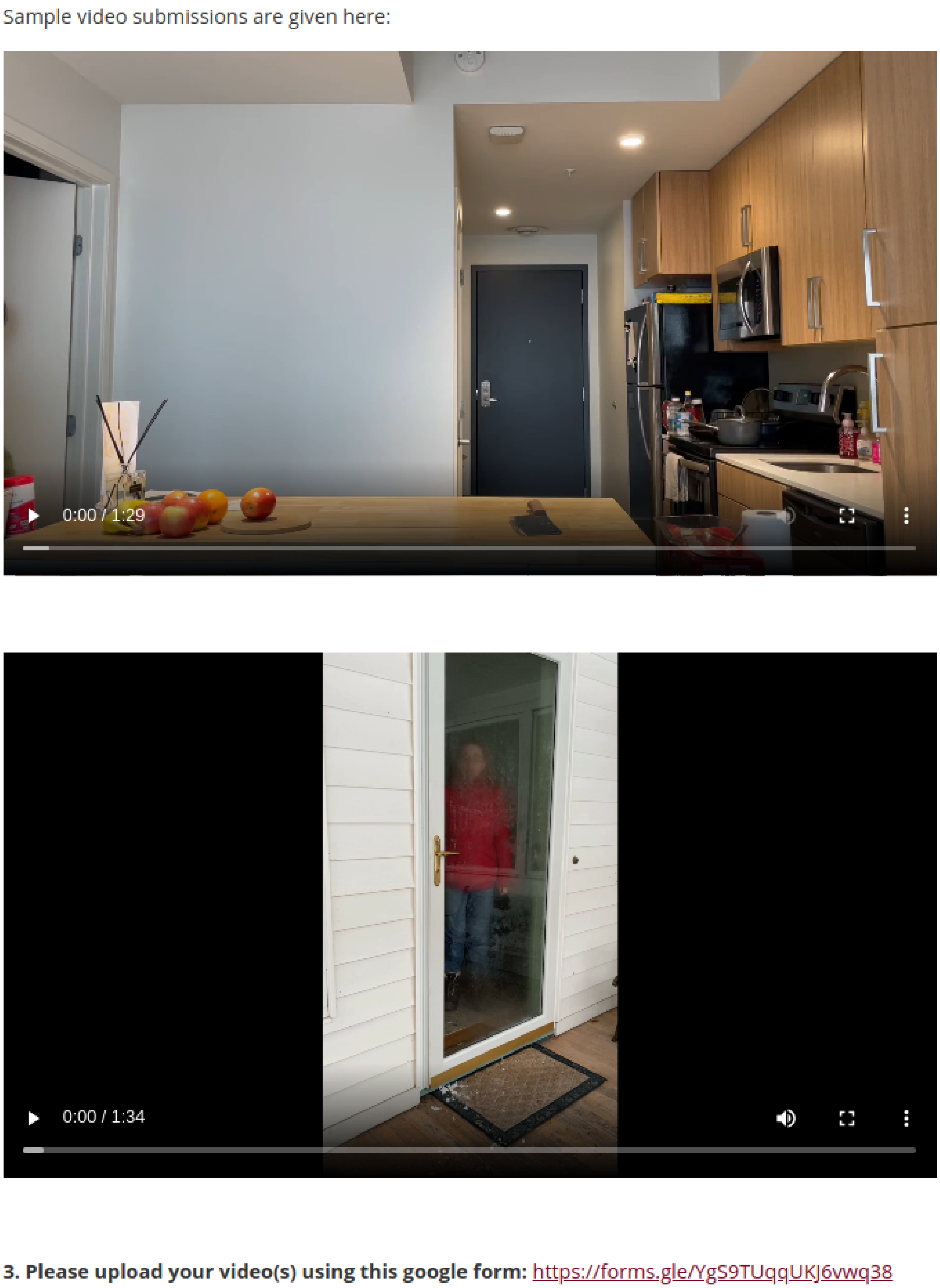
Please use the above images as reference for each of the four classes of gestures. You may make these gestures in any orientation.

When capturing your video, please do not use any filters, or add color-correction afterwards. Likewise, please do not edit the video after capture, unless it is necessary to adjust the brightness or crop. Do not alter the shape or orientation of the gesture or make any other edits.

We request that your video submission not contain geo-location data or audio. If included, these components of the video will be removed and the original video destroyed.

Sample video submissions are given here:

Figure A.4: Website part 4.

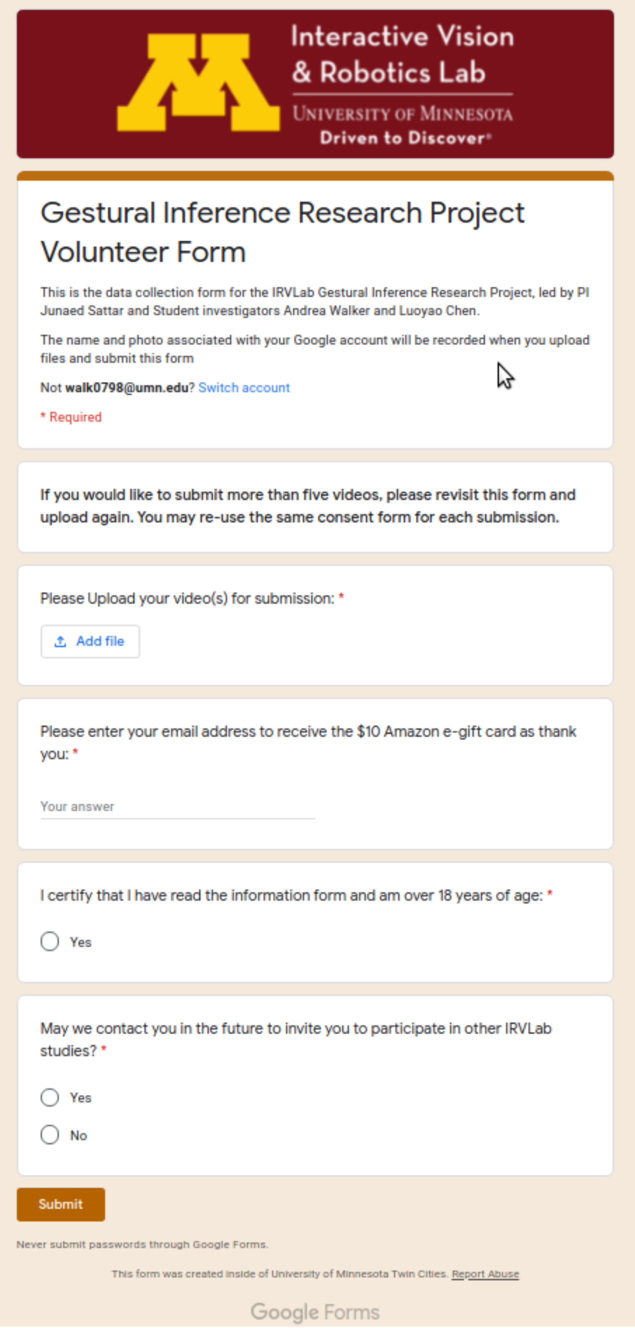


Sample video submissions are given here:

3. Please upload your video(s) using this google form: https://forms.gle/YgS9TUqqUKJ6vwq38

Figure A.5: Website part 5.

A.7 Intake Form



Interactive Vision & Robotics Lab
University of Minnesota
Driven to Discover®

Gestural Inference Research Project Volunteer Form

This is the data collection form for the IRVLab Gestural Inference Research Project, led by PI Junaed Sattar and Student investigators Andrea Walker and Luoyao Chen.

The name and photo associated with your Google account will be recorded when you upload files and submit this form

Not walk0798@umn.edu? Switch account

* Required

If you would like to submit more than five videos, please revisit this form and upload again. You may re-use the same consent form for each submission.

Please Upload your video(s) for submission: *

Add file

Please enter your email address to receive the $10 Amazon e-gift card as thank you: *

Your answer

I certify that I have read the information form and am over 18 years of age: *

○ Yes

May we contact you in the future to invite you to participate in other IRVLab studies? *

○ Yes

○ No

Submit

Never submit passwords through Google Forms.

This form was created inside of University of Minnesota Twin Cities. Report Abuse

Google Forms

Figure A.6: Intake form for research study