UNIVERSITY OF CALIFORNIA
RIVERSIDE

Scene Understanding for Autonomous Robotic Systems
Inferring Physical Interaction
Application to Assistive Robotics & Precision Agriculture

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Amel Dechemi

September 2023

Dissertation Committee:

Dr. Konstantinos Karydis, Chairperson
Dr. Amit K. Roy-Chowdhury
Dr. Salman Asif

The Dissertation of Amel Dechemi is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

As I reach the end of my academic journey, I would like to extend my heartfelt acknowledgment to those who have played a role in shaping this significant chapter in my life, and who have supported me in ways both seen and unseen. It is often said that it takes a village to have a PhD, and I consider myself fortunate to have been supported by not just one, but two invaluable villages.

The first one was at UCR- the professors, staff members, and fellow students. Your support and camaraderie during this phase have been immeasurable. I would like to express my appreciation to Dr. Konstantinos Karydis for the opportunity to undertake this academic journey. Your guidance has truly allowed me to develop a strong sense of reliance and determination, and explore avenues I might not have otherwise. In many ways, this experience has contributed to my personal and academic growth. I want to thank Dr. Amit K. Roy-Chowdhury and Dr. Salman Asif for accepting to be a part of my committee and their thorough evaluation of my dissertation. I also want to thank all the professors whose courses challenged me, broadened my knowledge, and improved my skills.

Over the years, I have had the honor of connecting, collaborating, and working alongside exceptional individuals. I would like to extend my sincere appreciation to my labmates at ARCS Lab: Georgia Kouvoutsakis, Pamodya Peiris, Dr. Lu Shi, Christopher Eng, Dimitris Chatziparaschis, Keran Ye, Hanzhe Teng, Joshua Chen, Tomás Olvera Hale, Merrick Campbell, Cody Simons, Dr. Zhichao Liu, Dr. Zhouyu Lu, Dr. Xinyue Kan, Xiao'ao Song, and Mehrnosh Ayazi for their contributions to my journey. I would also like to extend my thanks to Dr. Caio Mucchiani and Dr. Mohamed Karim Bouafoura for their

insights. Furthermore, I will express my appreciation to the PRT Lab for our collaborative works. Specially, I would like to extend my heartfelt gratitude to Georgia Kouvoutsakis, Pamodya Peiris, Chanisa Tangtananusak, and Dr Lu Shi for their support, memories and adventures.

Additionally, I thank the department of Electrical and Computer Engineering for for their exceptional support, particularly during the pandemic. Especially, Kim Underhill, who was a true lifesaver helping me navigating UCR's system.

Lastly, I want to express my deepest gratitude to my dearest village: my family. Your unwavering support, encouragement, and love have been the bedrock of my journey. Through every challenge and triumph, you have been there, providing me with strength, guidance, and a sense of belonging. Your belief in me has fueled my determination and resilience, and I am profoundly grateful for the sacrifices you have made to help me reach this point. Your presence in my life is a constant reminder of what truly matters, and I am blessed to have you by my side. Thank you for being my pillars of strength and my greatest source of inspiration.

This work has received financial support from USDA-NIFA under grant # 2021-67022-33453 and a UC MRPI Award. This dissertation has also yielded several master's thesis and undergraduate research projects involving students who made contributions to different facets of my research. I would like to acknowledge the participation of Vikarn Bhakri, Arjun Modi, Julya Mestas, Dannya Enriquez Barrundia, Pamodya Peiris, Merrick Campbell, Tomás Olvera Hale, and Christopher Eng. Segments of this work have previously appeared in the 30^{th} IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) and the 35^{th} IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

To my parents, Yamina & Abdelkader,

for the strength, guidance, and a sense of belonging.

To my siblings, Sarah, Dina, Youcef and Zino,

for the cheerful spirit & legendary humor.

To my nephews and nieces, Rita, Neil, Adam, and Anya,

for the priceless love & hugs.

ABSTRACT OF THE DISSERTATION

Scene Understanding for Autonomous Robotic Systems
Inferring Physical Interaction
Application to Assistive Robotics & Precision Agriculture

by

Amel Dechemi

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, September 2023
Dr. Konstantinos Karydis, Chairperson

In the context of computer vision and artificial intelligence, scene understanding is a complex task that involves a combination of image analysis, pattern recognition, context modeling, and often relies on advanced machine learning techniques. Scene understanding has applications in various fields, including autonomous navigation, robotics, surveillance, augmented reality, and more, where a deeper understanding of visual scenes is crucial for making informed decisions and interacting effectively with the environment. This dissertation aims to contribute to two applications: assistive robotics and precision agriculture. The first part of the dissertation focuses on the emerging research area of infant action recognition, specifically targeting the task of reaching—a significant developmental milestone. Existing action recognition models primarily cater to adults, leaving a gap in pediatric applications. To bridge this gap, BabyNet, a lightweight network, is introduced. It employs annotated bounding boxes to capture spatial and temporal relationships, accurately detecting reaching onset, offset, and complete actions. However, challenges emerge

due to the dataset's limited perspectives and the reliance on the detector network's performance. To overcome these limitations, E-BabyNet is proposed. It employs LSTM and Bidirectional LSTM layers to assess reaching actions and deliver precise onset and offset keyframes, handling transitions between actions. The second part focuses on the automation of leaf sampling for phenotyping known to provide highly accurate and timely information to growers. The first step in such automated procedure is identifying and localizing a leaf. For this purpose, we present a novel approach for leaf detection and localization using 3D point cloud with the ability to adapt to various designs. The preliminary results highlight successful indoor and outdoor leaf detection and localization. To better assess our approach, an actuation-perception framework was integrated with two distinct end-effectors, electrical- and pneumatic-based. Both end-effectors underwent evaluation within an indoor environment. Based on the obtained results, further experiments were conducted in the field, exclusively employing the pneumatic end-effector.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human visual system stands out for its remarkable ability to swiftly and accurately interpret the intricate visual world that surrounds us [48]. This incredible feat is commonly referred to as "*visual recognition*". In recent years, the field of computer vision has witnessed notable advancements, particularly in tasks involving the classification and precise identification of individual objects or object categories present within images. Furthermore, there has been an expansion of diverse complex tasks stemming from scene understanding, which has contributed to its broader scope. Key tasks include:

- **Object Detection and Recognition** involves identifying and localizing multiple objects within an image or video frame. This task goes beyond just labeling objects; it also provides information about where they are located in the image. Various algorithms, including convolutional neural networks (CNNs) and region-based approaches such as Faster R-CNN [55], YOLO (You Only Look Once) [128], and SSD (Single Shot MultiBox Detector) [101], are used to accomplish object detection.

- **Semantic Segmentation** aims to label each pixel in an image with a class label, thereby creating a pixel-wise segmentation map that highlights the different objects or regions in the scene. Fully convolutional neural networks (FCNs) [104], U-Net [130], and DeepLab [26] are popular architectures for semantic segmentation.

- **Object Tracking** involves following the movement of objects across consecutive frames in a video. It is crucial for maintaining the identity of objects as they move, allowing for analysis of their trajectories and interactions. Tracking algorithms use methods such as correlation filters [100], Kalman filters [94], and more advanced techniques such as Siamese networks [13] for tracking.

- **Pose Estimation** involves the task of determining the 3D positions and orientations of objects or entities within a scene. This process is crucial to comprehend the spatial relationships between objects and their relative poses. When it comes to human interactions and activities, the scene is analyzed to estimate the posture of the body and the positions of its joints.

- **Spatial Relationships** comprise understanding how objects are located in relation to each other and their surroundings, in terms of distance, direction, orientation, and proximity, and thus help in recognizing patterns and structures within a scene.

- **Contextual Understanding** involves considering the broader context of a scene to interpret objects accurately, and so analyzing the relationships, interactions, and associations between objects, as well as understanding how the scene's overall environment impacts the interpretation of individual elements.

- **Action Recognition** considers the task of identifying and categorizing the actions or activities taking place within a visual scene, often involving human or object movements. It aims to understand the dynamic aspects of a scene by analyzing the changes and interactions occurring over time. Widely adopted algorithms includes bidirectional Long Short-Term Memory (Bi-LSTM), 3D CNN [63], 3D Residual Networks (3D ResNet) [62], and Temporal Segment Networks (TSN) [163].

- **Scene Understanding in 3D** extends scene understanding into three dimensions by incorporating depth information. This can be obtained from depth sensors such as LiDAR or stereo cameras. It allows for a more accurate understanding of object positions in 3D space and can help address challenges such as object occlusion.

In the context of scene understanding, these tasks often complement each other to provide a comprehensive interpretation of a scene. Therefore, scene understanding plays a pivotal role in enhancing the capabilities of autonomous robotic systems, enabling them to make well-informed decisions and interact seamlessly with their surroundings. By comprehending the visual context of their environment, these robotic systems can analyze complex scenes, identify objects, assess spatial relationships, and recognize dynamic actions. Additionally, scene understanding equips these systems with the ability to interact with objects and humans intelligently, leading to safer and more efficient interactions–whether it is in industrial settings [45], agricultural operations [164], or healthcare [173]. In this dissertation, we present scene understanding approaches to aid robotic systems in inferring appropriate interactions within their surroundings, focusing on two principal applications: assistive robotics and precision agriculture.

Regarding the assistive robotics application, the approach encompasses **object detection and recognition, spatial relationships, and action recognition**. We develop a lightweight data-driven framework for video-based infant reaching action recognition to implement in closing action-perception loops of wearable robotic exoskeletons [84] for upper extremity pediatric rehabilitation. Furthermore, we extended our approach to address broader challenges and achieve more refined results. On the other hand, the precision agriculture-oriented approach involves **object detection and recognition, spatial relationships, and pose estimation in 3D**. The contribution lies in the development of an adaptive visual perception approach capable of accurately detecting and estimating the poses of leaves on tree crops, with the specific goal of aiding robotic plant phenotyping endeavors. To validate the effectiveness of our approach, we integrate and evaluate an actuation-perception framework utilizing two newly designed end-effectors, electrical- and pneumatic-based, within indoor settings. Building upon the outcomes from the indoor experiments, we further extend our experimentation to real-world field conditions, employing the pneumatic-based end-effector.

1.1 Related Works

1.1.1 Scene Understanding for Assistive Robotics

Scene understanding empowers assistive robotic systems to interpret and interact effectively with their environment. Through advanced computer vision techniques like object recognition and activity inference, they can recognize objects, comprehend gestures, and navigate complex spaces. Human action recognition stands as a crucial aspect of de-

veloping assistive robotics paradigm, enabling them to comprehend human activities and tailor their responses accordingly.

State-of-the-art video-based human action recognition algorithms have been steadily shifting from the use of Support Vector Machines or Hidden Markov models (e.g., [59,137]) to the use of deep learning networks, primarily due to the accessibility, adaptability, accuracy and decrease in time execution the latter can offer (e.g., [40,49,144,162]). One of the first significant attempts [144] used two separated convolutional neural networks (CNNs) trained to extract features from a sampled RGB video frame paired with the surrounding stack of optical flow images. As optical flow is computationally expensive and has a burdensome optimization process, most follow-on works used a CNN to learn the optical flow prediction [46,72], thus reducing the number of parameters as only one network is needed. Other methods explored the advantages of LSTM structures [40] to incorporate motion by updating the pooling of the features across time [35,53].

Despite some remarkable results achieved by RGB-video-based methods to date, some key challenges remain. These include, for example, background clutter, illumination disparity, pose/viewpoint variation, to name a few. One way to improve recognition performance under these challenges is via skeleton data representations that do not contain color information. Early relevant works did not use information regarding internal dependencies between body joints [50,99,158,176]; more recent works apply graph convolutional networks to extract features by building a skeleton graph composed of vertices and edges to represent joints and links, respectively [151,174,175]. These approaches rely on datasets that contain mostly motion actions performed by adults [23,31,58,90,142,145].

However, the transition to the realm of young children and infants introduces a distinct set of challenges. Identifying human motor actions that emerge early in life (e.g., spontaneous movements of arms and legs, kicking, crawling, etc.) is an emerging vision-based action recognition research direction [24,42,122]. The ability to *automate* the process of detecting, recognizing, and classifying actions performed by young children and infants from visual data can prove useful in many pediatric applications. Examples include monitoring for safety [57, 119, 150], studying infants' interaction with caregivers [83], identifying markers for diagnosis of neuromotor disorders [24, 42, 64, 146, 167], and closing action-perception loops for smart environments and automated assistive devices [41, 85, 106, 154], as well as autonomous wearable robots for pediatric rehabilitation [84] whereby vision-based action recognition can help guide autonomous reasoning regarding the amount of passive/active feedback to provide to the user.

The challenge is the inherent movement variability within and across young humans as a natural result of learning and growth [51]. Infants' bodies have distinct proportions and can attain distinct poses compared to those of adults [140]. For example, the kinematic properties of infant reaching (the motor action of interest in this work) changes over time as infants learn how to adapt to environmental, task, and biomechanical constraints [89]. In fact it may take years for children to achieve smooth and straight reaching trajectories similar to those seen in adults [12, 136]. Therefore, existing skeletal models or learning-based pose estimation methods often used in adult action recognition may not be the best fit for pediatric action recognition [41, 150].

Still, an interesting shift toward pediatric rehabilitation paradigms that involve assistive robots is observed. The NAOTherapist [124] allows for non-contact upper-limb rehabilitation autonomously, by utilizing a social robot that performs a set of prescribed arm-poses for a child to imitate. Movements here are captured by a Kinect depth camera, stored as 3D skeletons and then compared with entries from a knowledge base. Other applications use multiple cameras to alleviate issues such as occlusion that is more likely to happen in infant rehabilitation sessions [41, 85, 122, 154]. Nevertheless, the number and type of infant motor actions that have been utilized in action recognition research still remains limited. Under these circumstances, developing algorithms that accurately and reliably recognize motor actions performed by infants is a challenging process.

1.1.2 Scene Understanding for Precision Agriculture

Precision agriculture refers to the application of advanced technologies to optimize, customize farming practices, and enables growers to make informed decisions that enhance resource efficiency, reduce waste, minimize environmental impact, and ultimately improve crop yields and quality. It involves using data, information technology, and specialized tools to make informed and precise decisions about farming practices.

In the realm of precision agriculture, scene understanding emerges as an impactful tool that unlock insights from visual data in fields for various tasks. Tasks within precision agriculture can be categorized based on the level of physical interaction required with the crops. For tasks like monitoring and inspection [20, 25, 116], or spraying [110], the combination of a precise field map and accurate robot positioning proves sufficient. These tasks rely on accurate data collection and application without the need for direct

physical contact with the crops. In contrast, interactive tasks such as pruning [16, 81, 105], harvesting [112, 171, 172], or sampling [2, 5, 19, 120] demand an additional layer of precision–specifically, an accurate pose estimation. This information is crucial to enable robots to approach the crops effectively and retrieve samples from it successfully. There is a notable emphasis on developing non-destructive techniques across the field of agriculture. These methods prioritize obtaining valuable information without causing damage or altering the integrity of the plant or tree. For instance, non-destructive techniques are increasingly employed for crop assessment, allowing researchers and growers to gather insights about plant health, growth, and quality without harming the crops [61,96,132,165]. These methods span from remote sensing and hyperspectral imaging to ultrasound and infrared thermography. Consequently, there is a growing interest in developing semantic and instance segmentation for agricultural applications.

Still, while remote sensing techniques provide valuable data on external features and certain internal characteristics, they fall short in directly detecting intricate internal processes. These traits include nutrient concentrations, metabolic activities, gene expressions, cellular changes, and identifies hidden toxicities and deficiencies with symptoms that cannot be observed solely through remote sensing. Moreover, remote sensing lacks the resolution to capture microscopic structures, and genetic variations within plants. In contrast, physical sampling and follow-on analysis of plant specimens (e.g., leaves, shoots, etc.) can offer reliable data regarding the internal state of the plant (e.g., leaf tissue analysis [98,135,177]).

The effectiveness of automating such physical sampling procedures relies heavily on the visual perception system's ability to provide precise and accurate information about

the target crop and relevant environmental context [80]. Most approaches have focused on fruit/vegetable targets by harnessing distinct colors and/or shapes [5,27,52,118,125]. Identifying a leaf and its stem on a tree is a more challenging problem with classical methods since the process is more likely to be affected by occluded and overlapped leaves as compared to segmenting fruits out of a canopy [5]. This presents similar yet unique challenges compared to fruit (and broader canopy) identification.

1.2 Objectives & Contributions

1.2.1 Infant Action Recognition

Approaches tailored to and validated with adults might not be the most suitable for implementation in the context of recognizing infant actions, owing to various factors. These include differences related to body properties and kinematic patterns [15]. Further, infants need exposure to complex (unconstrained) and variable environments to learn [86], as well as immediate rewards and constant motivation to perform motor tasks in these environments [44]. Consequently, training and assessment approaches should be applied in such environmental contexts [93]. Our contributions are in line with this approach through the analysis of infant reaching actions in unconstrained environments.

First, we develop, in two phases, a new annotated dataset that includes diverse reaches performed while in a sitting posture by different infants in unconstrained environments (e.g., in home settings). Next, we introduce our approach, *BabyNet*, that uses the spatial and temporal connection of annotated bounding boxes to interpret onset and offset of reaching, and to detect a complete reaching motion. We evaluate the efficiency of

our proposed approach and compare its performance against other learning-based network structures in terms of capability of capturing temporal inter-dependencies and accuracy of detection of reaching onset and offset. Building upon the obtained outcomes, we extended the structure to enhance the action recognition by including bimanual reaching detection and reducing false detection, resulting in the development of *E-BabyNet*. The structure consists of two main layers based on two LSTM and a Bidirectional LSTM (BiLSTM) model, respectively. The first layer provides a pre-evaluation of the reaching action for each hand by providing onset and offset keyframes. Then, the biLSTM model merges the previous outputs to deliver a final outcome of the reaching actions detection for each frame including the reaching hand. We evaluate our approach against four other lightweight structures using a fully annotated dataset comprising 375 infant reaching actions performed in sitting positions by different subjects [38].

1.2.2 Object Detection & Pose Estimation for Robotic Plant Phenotyping

Physical sampling of leaf specimens is a critical component to help assess the plant's overall health and attributes that remote sensing might omit. The contribution aims to enable spatio-temporally dense sampling of leaves to support plant phenotyping research using mobile agricultural robots. For this purpose, we introduce an approach for detecting and localizing candidate leaves using 3D point cloud data from a depth camera. The detection and pose estimation are validated through indoor experiments with a real avocado tree. Subsequently, the visual perception approach is integrated into two separate end-effectors demonstrating promising outcomes in indoor settings. The results demonstrate that the pneumatic-based end-effector can handle more clustered areas than electrical-based

one. The latter encounter challenges in producing consistent results. Following this, we expand our experimentation to encompass outdoor field conditions across diverse scenarios employing the pneumatic-based end-effector. The assessment encompass both static random sampling and planned sampling tasks The overall experimental testing demonstrates that our proposed approach can enable our mobile manipulator and custom end-effector systems to successfully detect, localize, and cut leaves. Moreover, it demonstrates the capability to adapt to the unique designs provided by each respective end-effector.

1.3 Dissertation Layout

The remainder of this dissertation is as follows: Chapter 2 focuses on **infant action recognition** within the field of assistive robotics. In this context, two novel action recognition approaches—namely, *BabyNet* and *E-BabyNet*—are introduced along with a newly developed dataset. Chapter 3 introduces an **object detection and pose estimation** approach for plant phenotyping. The visual perception strategy is constructed and integrated with two newly designed end-effectors. Lastly, Chapter 4 summarizes the dissertation while outlining forthcoming potential works.

Chapter 2

Infant Reaching Action Recognition in The Wild for Assistive Robotics

In this chapter, we present two novel approaches for infant action recognition with the objective of providing a light-weight (based on the number of trainable parameters) structure of comparable efficiency with (significantly) larger ones. For this purpose, we introduce *BabyNet*, a new network structure built upon a long short-term memory (LSTM) module to model different stages of reaching action through a spatial-temporal interpretation. The results shows that the structure can challenge the performance of significantly larger structures by 66.27% average testing accuracy and is able to detect reaching action with a precision and recall scores of 0.66 and 0.49, respectively. Despite the performance, *BabyNet* is challenged by the lack of viewpoint of the training dataset and the transition

between the reaching and no reaching stages. In addition, if a reaching action is confirmed, the reaching hand is not specified which can be essential information for assessing the infant motion and/or closing the action-perception loop in an upper extremity pediatric wearable robotic device.

With the aim of addressing these limitations, we propose *E-BabyNet*, a new two-layered learning structure able to assess separately right-handed and left-handed reaching action, as well as to handle bimanual reaching. In addition, we integrate a Bidirectional LSTM (biLSTM) structure, which allows for better identification of the transition between the no-reaching and the reaching phases. Results illustrate the effectiveness of our approach and ability to provide reliable reaching action detection and offer onset and offset keyframes with a precision of average of one frame. Moreover, the biLSTM layer handles the transition between reaching actions and reduced false detections.

To assess our approaches, a novel dataset focusing on infant reaching was developed in two phases. During the*BabyNet*'s development phase, a collection of 193 reaching instances executed by 21 infants was gathered from 20 videos. In order to enhance the generalizability of reaching actions, we expanded the dataset to encompass 375 instances of reaching, involving 40 different subjects, thereby enabling the testing of our enhanced structure *E-BabyNet*. Annotations and bounding boxes that describe reaching properties (e.g., reaching onset/offset, object touched, etc.) are also included.

The chapter is structured as follows: Section 2.1 discusses related works, while Section 2.2 presents the developed dataset. In Section 2.3, we delve into the details of *BabyNet*, including its structure, implementation, and the results and discussion. Next,

Section 2.4 covers *E-BabyNet*, following the same structure as for *BabyNet*. Lastly, we conclude the chapter with Section 2.5.

2.1 Related Works

The ability to employ machine vision and artificial intelligence to identify, comprehend and anticipate various human activities is critical in order to develop effective and interactive human-computer interfaces [166], and healthcare systems [8, 36, 54]. Such an ability falls within the realm of developing vision-based human action recognition algorithms (e.g., [23,31,58,90,142,145,149,178]), often using different sensing modalities. Most existing works have considered the development of action recognition algorithms based on datasets that contain (young) adult motions (e.g., [30, 75, 99, 159]) considering the large possible actions and application domains afforded by such populations.

We focus on the emerging research thrust of *infant action recognition*, mainly toward pediatric applications. Some examples include identification of early signs of neuromotor disorders [3, 134] and assessment of the performance of behavioral and physical therapy through smart environments and assistive wearable robotic devices [85, 122]. Unlike datasets containing (young) adult activity, datasets containing infant motions are sparse due to regulations protecting children's privacy being stricter as infants have no control over the release of the data, which in turn has hindered the development of action recognition methods for this population.

There has been a keen effort in the research community to address the current shortage of infant activity datasets [22, 37, 67, 71, 122]. Such datasets contain information

retrieved as RGB images, depth images, or 2D/3D skeletons through one or multiple cameras, and were collected following either passive or interactive paradigms to elicit infant motion. Motivated by an observed decrease in the accuracy of pose estimation approaches trained on adult data, Sciortino et al. [141] developed a dataset collected using publicly available videos of 104 children in natural environments and manually annotated 22 body keypoints. Another significant dataset is the Moving INfants In RGB-D (MINI-RGBD) based on the Skinned Multi-Infant Linear body model (SMIL) [67]. The authors used realistic shapes and textures to produce RGB and depth images for 2D and 3D joint positions by mapping real infants' movements to the SMIL model. Recently, the InfAct (Infant Action) dataset was introduced in [71]. The dataset contains 200 video clips of infant activities and 400 images of infant postures. It also includes structured action and transition segmentation labels.

The specific action of interest in this work is reaching. Research has shown that main developmental milestones are gradually achieved during the first two years of an infant's life [4, 88]. Reaching is among the most significant milestones since the ability of infants to explore and interact with their environment directly impacts their motor, social, perceptual, and cognitive development [4, 14, 102, 103, 169]. In this context, automating the identification of such motion is essential and requires approaches able to accurately and consistently recognize the infant's movements. As it pertains to reaching tasks, specifically, datasets from (young) adults are inappropriate to be used for infant action recognition algorithm development. This is due to the intricacy and the variation of an infant's reaching motion which evolves uniquely for each infant and could be shaped by the location, the size,

and the shape of the object [12,86,87,97]. Besides, the straightness and smoothness of infant reaching dramatically increase throughout the first two years whilst speed and jerk decrease; these can server as key indicators to determine attainment of full level of control [44,152,160]. However, experienced reachers tend to have much less variable and consistent reaching actions, thereby fundamentally differing from reaching actions of developing humans [12].

This work has a twofold aim. First, to develop a new dataset focusing on infant reaching. Second, to develop a machine learning algorithm for infant reaching action recognition. The dataset is constructed based on diverse online-shared videos that demonstrate reaching actions of both typically-developing infants as well as infants with arm mobility challenges (all between 6–12 months of age). Annotations and bounding boxes that describe reaching properties (e.g., reaching onset/offset, object touched, etc.) are also included. Next, two new network structures aimed at infant reaching action recognition were developed. The first network, *BabyNet*, is built upon a long short-term memory (LSTM) module to model different stages of reaching action through a spatial-temporal interpretation. Our motivation is to provide a light-weight (based on the number of trainable parameters) structure of comparable efficiency with larger ones. The second structure, *E-BabyNet*, is a two-layered learning structure able to assess separately right-handed and left-handed reaching action, while effectively handling bimanual reaching which was difficult to detect by the previous network.

2.2 Dataset of Infant Action Recognition

The dataset is based on videos collected from the YouTube^{TM} online video-sharing platform [1] using search terms such as 'infant', 'reaching', 'grabbing', and 'sitting'. Videos were included in the dataset if they displayed awake and behaving infants: (1) up to 12 months of age, (2) placed in a sitting position, (3) reaching for objects (regardless of shape and/or position), (4) performing at least one reaching action during which the camera remained stationary or moved slightly, and (5) performing at least one reaching action during which both hand and object were visible. Both typically-developing infants and infants with arm mobility challenges were considered. The majority of videos were recorded in natural (unconstrained) environments (e.g., family's home, clinic). The clothes of infants, presented objects, and the background varied from video to video. Figure 2.1 shows some illustrative samples.

Figure 2.1: Samples frames from the infant reaching dataset employed in this work.

[1] Video handling procedures were according to YouTube's statement on fair use of videos for research purposes.

To obtain the reaches, we performed a manual annotation process. For this reason, we defined the reaching action by two phases: Onset (RN) and Offset (RF). The onset is defined as the first frame in which movement of the infant's hand (left, right, or both) toward an object presented to the infant is initiated. Associated with RN is the Onset Keyframe (K_{RN}). Similarly, the offset is defined as the first frame in which the infant's hand(s) make contact with the object; we also define the Offset Keyframe (K_{RF}). Examples of onset and offset keyframes are shown in Figure 2.2. Note that in our definition the offset can be triggered either by the full hand grasping the object or as an intentional contact by parts of the hand such as the fingers or the palm.

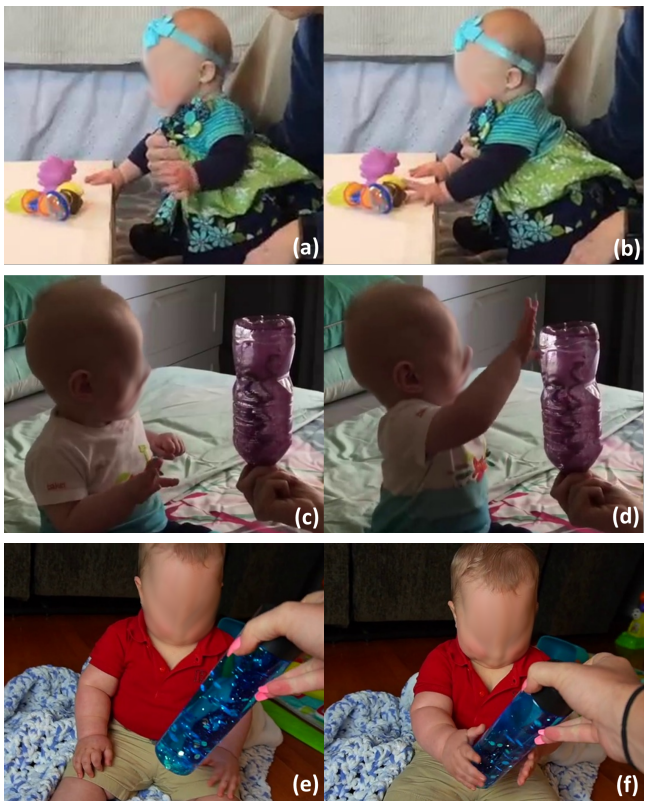

Figure 2.2: Samples annotated Reaching Onset (left panels) and Offset (right panels) frames. Panels (a)-(b): left-handed reach, panels (c)-(d): right-handed reach, Panels (e)-(f): bimanual reach.

Identification and collection of appropriate videos took place in two phases. In the first phase, a total of 193 reaches performed by 21 distinct subjects were collected through 20 videos (Table 2.1). Out of the 21 subjects, five subjects had a medical diagnosis (four with Down syndrome and one with congenital anomaly; information was provided in the video description). In most cases, the video description was detailing the age and gender of the subject; in a few exceptions that this information was omitted (entries marked with an * in the tables), our research team empirically estimated these characteristics. 607 images were randomly sampled out of total of 2,984 frames and annotated. For each frame, we created bounding boxes of the infant, their left hand, their right hand, and the objects involved in the reaching action. Annotation of the selected 607 frames resulted in a total of 3,194 bounding boxes.

During the second phase, we expanded the dataset, acquiring a total of 375 reaching instances performed by 40 distinct infants, all within the age range of up to 12 months. (see Table 2.2). Among all the subjects, seven were reported in the videos' description as neuro-divergent with five diagnosed with Down syndrome, one with congenital anomaly, and one with autism spectrum disorder.

The distributions of reaching actions are shown in Figure 2.3 and Figure 2.5. For the left hand, 59.24% of the reaching actions are within a range of 3 to 15 frames. Note that 34.41% of the left-hand reaches last only between 3 to 7 frames. Longer reaches for the left hand case are considerably fewer. Indeed, reaches lasting between 31 to 48 frames and 49 to 63 frames comprise only 6.37% and 1.27% of the total number of left-hand reaches, respectively. Similarly, the right-handed reaching actions have about the same percentage

Table 2.1: Subjects and Annotated Reaches per Subject During *BabyNet*'s Development

Subject ID	Age [months]	Gender [M / F]	# Reaches LH	RH	Total	Total Frames
T01	6-8	F	5	3	8	205
T02	8-10	M	8	4	12	217
T03	11-12*	M	2	8	10	55
T04	6-12	F	3	5	8	222
T05	10-12	M*	3	1	4	94
T06	10-12	M*	0	2	2	63
T07	6-7*	M	2	3	5	154
T08	<12*	M	4	5	9	285
T09	6	F	3	6	9	165
T10	6	F	3	3	6	57
T11	6-8	M	2	1	3	41
T12	8	F	16	33	49	475
T13	10	F	5	9	14	84
T14	6	F	1	1	2	55
T15	9	M	20	20	40	497
T16	7	F	2	1	3	60
D01	6-9	F	1	-	1	23
D02	10	M	-	1	1	31
D03	<12*	M	4	1	5	57
D04	9-12	M	1	-	1	12
D05	<12*	M	1	-	1	7
Total	–	–	**86**	**107**	**193**	**2843**

of actions in the range of 3 to 15 frames as in the left hand case. Further, the number of left- and right-hand reaches is roughly the same for actions within the range of 3 to 7 frames, and for actions within the range of 8 to 15 frames. Actions with a length exceeding 30 frames represent up to 10% of the total number of right-hand reaches.

Table 2.2: Subjects and Annotated Reaches per Subject During *E-BabyNet*'s Development

Subject ID	Age [months]	Gender [M / F]	# Reaches LH	RH	Total	Total Frames
T01	6-8	F	5	3	8	205
T02	8-10	M	8	4	12	217
T03	11-12*	M	2	7	9	103
T04	6-12	F	3	5	8	222
T05	10-12	M*	3	-	3	84
T06	10-12	M*	-	1	1	52
T07	6-7*	M	-	2	2	73
T08	<12*	M	4	5	9	285
T09	6	F	3	6	9	165
T10	6	F	3	3	6	57
T11	6-8	M	2	1	3	41
T12	8	F	18	23	41	395
T13	10	F	5	8	13	220
T14	9	F	-	1	1	22
T15	12	F	-	1	1	24
T16	6	F	1	1	2	55
T17	7	M	15	5	20	287
T18	9	F	4	3	7	88
T19	7	F	4	10	14	134
T20	9	M	16	20	36	453
...	...	...	...	...	...	...

Subject ID	Age [months]	Gender [M / F]	# Reaches LH	RH	Total	Total Frames
T21	6	F	5	2	7	107
T22	6	M	10	19	29	383
T23	7	F	6	17	23	459
T24	8	F	1	19	20	394
T25	7	M	3	4	7	123
T26	3-6	F	3	3	6	107
T27	8	M	1	5	6	53
T28	7	F	1	1	2	33
T29	5-7	M	3	10	13	223
T30	9-10*	F	11	16	27	404
T31	8	M	-	1	1	34
T32	5	M	1	1	2	78
T33	7	F	1	-	1	21
D01	6-9	F	1	-	1	23
D02	10	M	-	1	1	31
D03	<12*	M	3	-	3	36
D04	9-12	M	1	-	1	12
D05	<12*	M	1	-	1	7
D06	12	F	2	-	2	50
D07	<12*	M	8	9	17	125
Total	–	–	**158**	**217**	**375**	**5865**

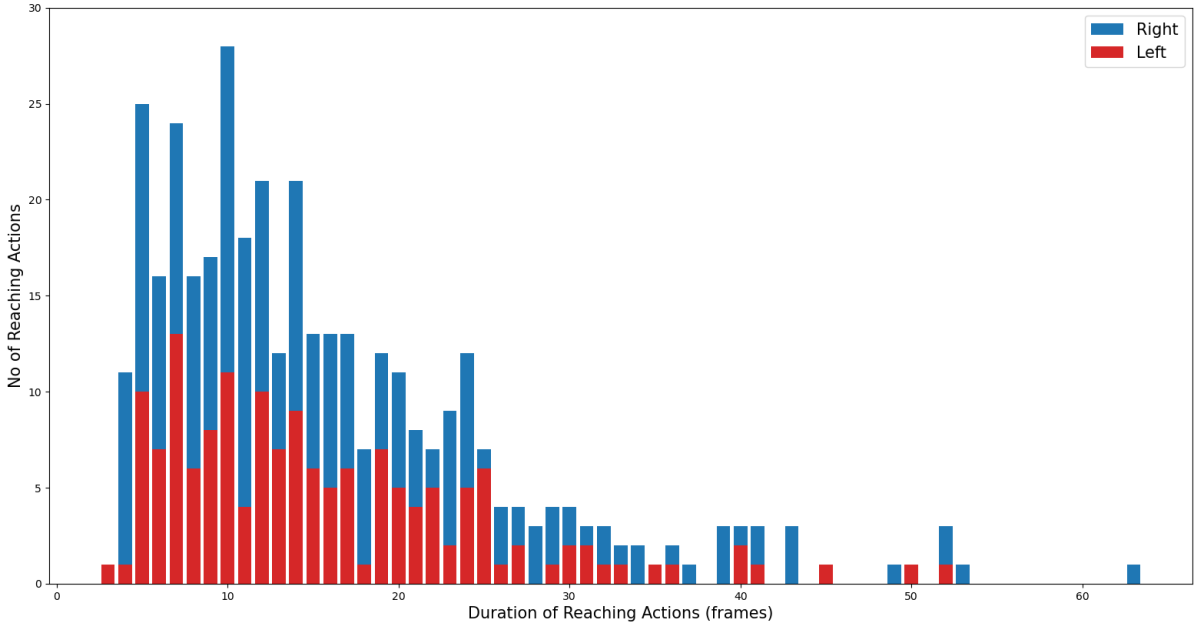

Figure 2.3: Distribution of duration of reaching actions performed by left and right hand.

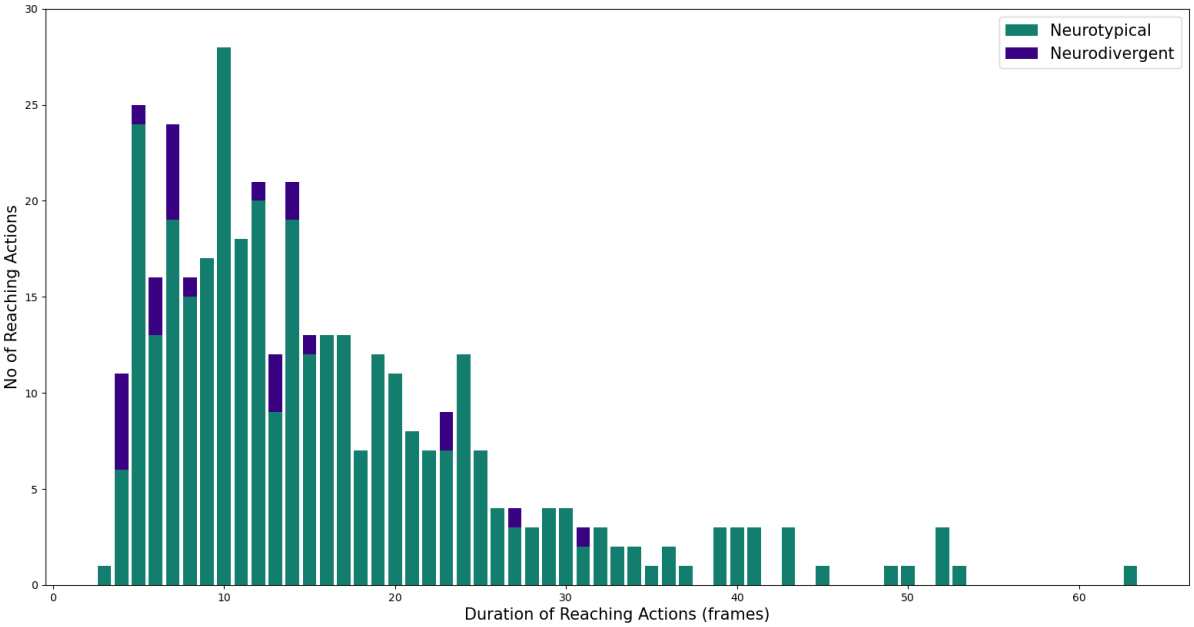

Figure 2.4: Distribution of reaching actions duration performed by neurotypical and neurodivergent infants.

Figure 2.4 depicts the reaches in terms of a neurotypical/neurodivergent subject split. It is worth noting that there was a total of 26 reaches from neurodivergent subjects, and out of these, 63.64% had a duration in the range of 3 to 7 frames. In contrast, 68.66% of the reaches performed by neurotypical subjects lasted for 8 to 15 frames.

Besides reaching annotations, we also developed bounding box annotations for every image extracted from the reaching actions. We manually annotated the left hand (LH) and the right hand (RH) of the subject as well as the presented objects (OBJ) for

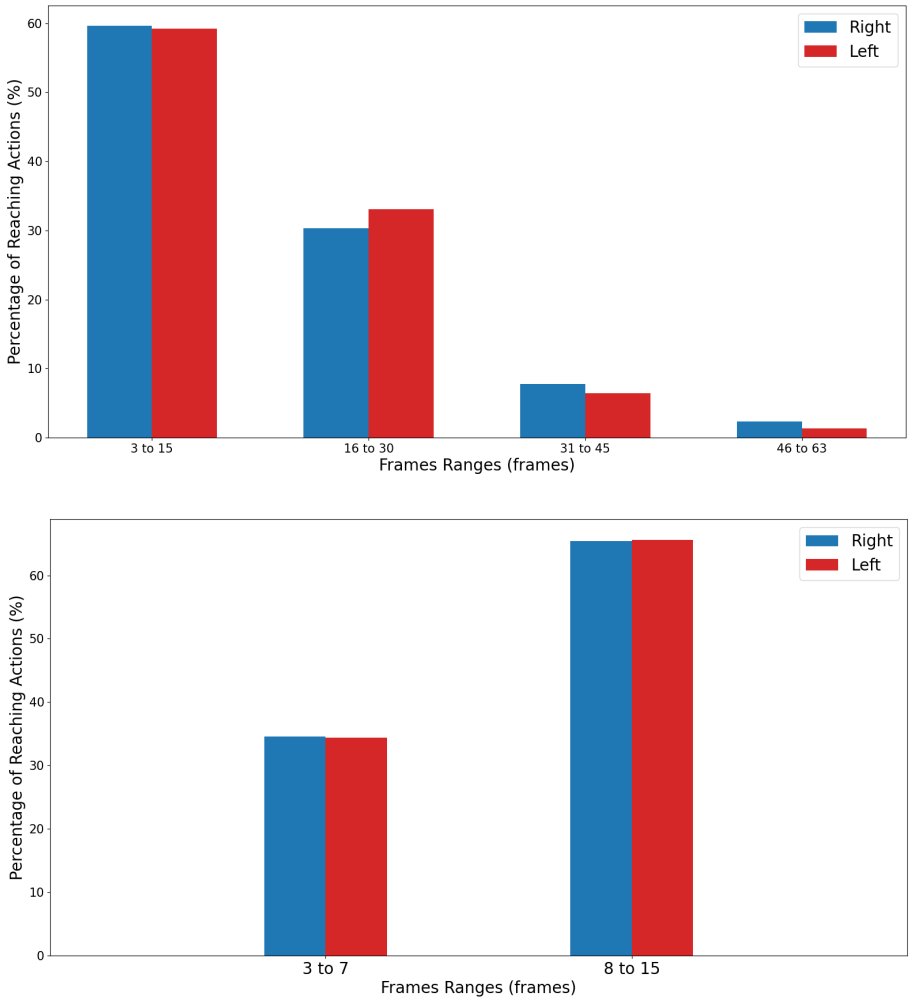

Figure 2.5: Comparative number of reaches between left and right hands.

each image throughout the duration of every reaching action. A total of 5865 images were annotated resulting in 16337 bounding boxes. These were employed to train, test, and assess the efficacy of our proposed action recognition approach as discussed next. In order to extrapolate the reaching action using the annotated dataset, we extract the distance between the center of the LH and RH bounding boxes with the object (OBJ) to be reached, and compute their intersection of union (IOU).

2.3 Approach I: *BabyNet*

2.3.1 Proposed Method

Let $\mathcal{B} = \{b_i^{\mathcal{F}}\}$ and $\mathcal{H} = \{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ denote the objects and the left and right hand detected in a sequence of frames $\mathcal{F}$, respectively. Each detection is associated with a bounding box with its center defined by x and y coordinates (in pixels); C_{b_i}, C_{h_L} and C_{h_R} denote the object, the left hand and the right hand bounding boxes, respectively. Let also $D_{b_i} := \{H_{b_i}, W_{b_i}\}$ denote the bounding box dimensions for objects. Likewise, let $D_{h_L} := \{H_{h_L}, W_{h_L}\}$ and $D_{h_R} := \{H_{h_R}, W_{h_R}\}$ denote the bounding box dimensions for the left hand and for right hand, respectively. These are defined at each frame $\mathcal{F}_j$. At initialization, both keyframes, K_{RN} and K_{RF}, are set as the first frame of $\mathcal{F}$. The obtained bounding box detections are used to correlate spatio-temporal patterns between $\mathcal{H}$ and $\mathcal{O}$, and undergo two separate processes for the Reaching Onset phase (RN) and the Reaching Offset phase (RF) as shown in Figure 2.6

2.3.1.1 The Reaching Onset Phase

First, we compute the distance d_j^i between each hand $\{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ and the object $b_i^{\mathcal{F}}$ in the current frame $\mathcal{F}_j$. Next, we evaluate $d_j^i - d_{j-1}^i$ as follows: if $d_j^i - d_{j-1}^i < 0$, the onset is confirmed and $\mathcal{F}_{RN}$ is kept as the onset keyframe K_{RN}; however, if $d_j^i - d_{j-1}^i \geq 0$ and continues increasing for four consecutive frames, the current onset is invalidated and the

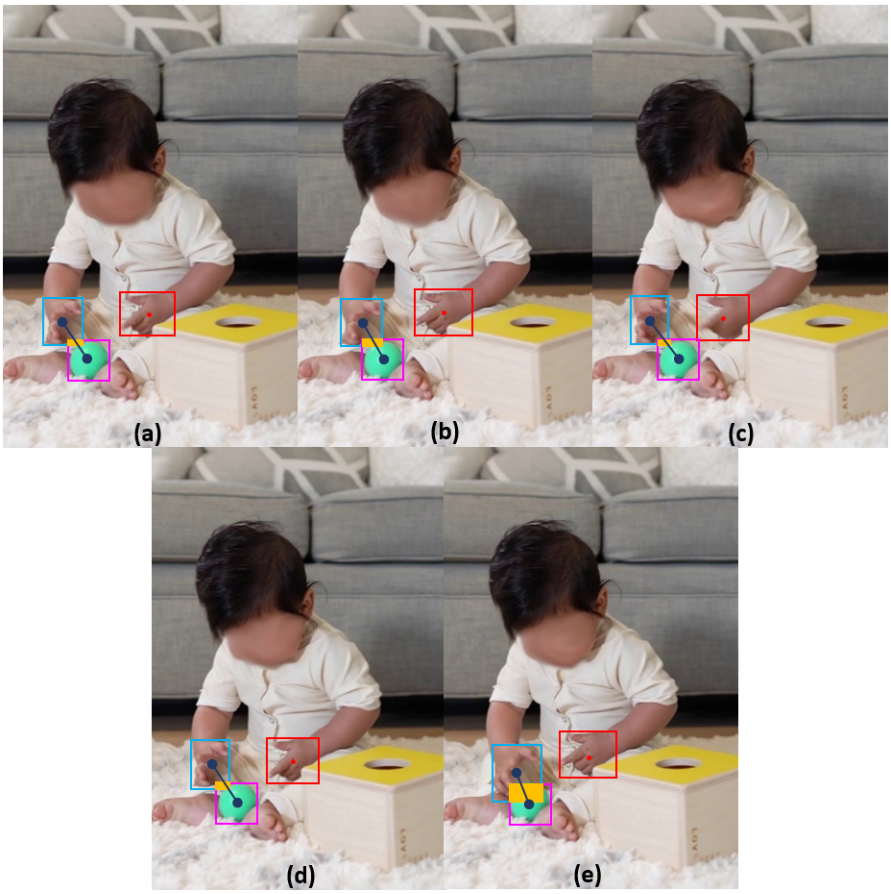

Figure 2.6: Qualitative illustration of our approach through an actual reaching action. (a) Onset: the movement of the right hand (cyan) is initiated to reach the object (magenta). During the (b)-(d) reach and (e) offset: the distance between the right hand and the object (dark blue) decreases as the intersection of union (orange) increases until the IOU value stabilizes (which indicates a successful reaching action). The distance between the left hand (red) and the object (magenta) does not decrease significantly and hence the respective IOU remains null which confirms that no reaching action is performed by the left hand.

onset keyframe K_{RN} is updated as $\mathcal{F}_j$. This is a practical way we can use to help prevent false detections considering the fact that the shortest reaching motion lasts for three frames.

2.3.1.2 The Reaching Offset Phase

To confirm the contact of the hand with the object, the Intersection of Union (IOU) is estimated between each hand $\{h_L^{\mathcal{F}}, h_R^{\mathcal{F}}\}$ and the object $b_i^{\mathcal{F}}$. The computed IOU is compared against a threshold value determined (a hyperparameter of our approach). This threshold can be estimated empirically via the data analysis process. The keyframe K_{RF} is updated as the current frame $\mathcal{F}_j$ as long as the IOU is less than the threshold and, thus no offset is confirmed. Otherwise, if contact is detected, the keyframe K_{RF} is definitively set as the current frame $\mathcal{F}_j$, and a new reaching action is initiated.

2.3.1.3 Core Structure of *BabyNet*

The proposed *BabyNet* uses the LSTM structure to learn the relation between the bounding boxes through the input consisting of the distances and intersection of union (IOU). The output is the scores for onset (RN), offset (RF), reach (R) and no reach (NoR) are used to update identified keyframes. The reach (R) is the label used for all frames between the onset RN and offset RF frames. Similarly, the no reach (NoR) is the label for all frames before the onset and after the offset. The flow chart of our proposed algorithm for infant reaching action recognition *BabyNet* is depicted in Figure 2.7.

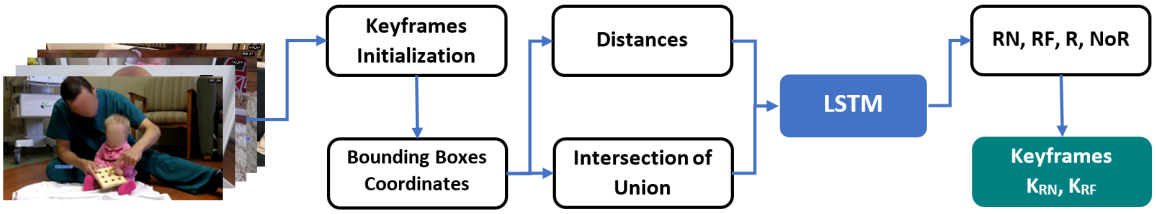

Figure 2.7: The underlying process followed by the *BabyNet* structure.

2.3.2 Implementation & Experiments

We considered four baseline network structures to serve as the basis to evaluate our proposed structure's performance.

- **Multi-Layer Perceptron (MLP)**. We trained a four-layer network with two inputs and four outputs.
- **ResNet**. Starting with a pretrained ResNet-50 model [66], the last Bottleneck block in the fourth layer of the network was retrained along with the fully connected layer in an effort to examine if overfitting can be avoided.
- **ResNet+LSTM**. An LSTM block was integrated after the average pooling layer of the final residual block of the aforementioned ResNet-50 model in an effort to examine if temporal correlation features of reaching actions can be captured.
- **LSTM with Optflow (O-LSTM)**. A single-layer LSTM with 50% dropout was trained to leverage the information provided by the optical flow images.

All models except for O-LSTM use RGB images of size 224x224 directly as inputs and were evaluated on our dataset and were trained with learning rate 0.001 using Adam optimizer and cross entropy loss. Both ResNet and ResNet+LSTM model were evaluated

with the data augmentation (D.A.) by altering randomly the images through techniques including shift, scale and rotation. The O-LSTM uses optical flow images obtained with the Farneback method [47] and trained with a learning rate of 0.0001 using Adam optimizer and cross entropy loss and inputs of a flattened image of size 1x12,288 (from a reduced-size 64x64 RGB source image) to reduce the training time.

BabyNet has two inputs and four outputs (RN, RF, NoR, R). Preliminary testing showed that selecting two inputs can improve temporal correlation predictions without overfitting. *BabyNet* was trained with learning rate 0.001 using Adam optimizer and cross entropy loss. In initial testing we used a small dataset of 63 reaches with a 60% training, 15% validation, and 25% testing split. While *BabyNet* (and MLP) can perform well with small datasets (as desired), larger structures (ResNet variants and O-LSTM) were found to overfit the dataset. To resolve this issue, we tested larger networks with a portion of the dataset comprising 193 reaches while testing *BabyNet* and MLP at 63 reaches.

2.3.3 Experimental Results

Comparison results, including classification accuracy (range: $[0 - 100\%]$) and precision/recall scores (range: $[0 - 1]$), are shown in Table 2.3. Including the precision/recall scores of both no reaches (NoR) and reaches (R) serves a dual purpose: (1) to examine the trade-off between different methods; and (2) to help attain a more clear distinction in terms of the structures' capability to correctly differentiate between no reaches and reaches.

We first compared the performance of the family of ResNet and ResNet+LSTM structures. The ResNet structure with data augmentation achieves the best performance with an average testing accuracy of 58.16%. With reference to Table 2.3, we can observe that

the models with data augmentation give nearly the same results as with those without data augmentation (cf. 58.16% to 53.43% in ResNet and 54.42% to 54.21% in ResNet+LSTM). To properly gauge the effect of data augmentation on accuracy, we compared the trade-off between precision and recall of the two networks. The ResNet model with data augmentation has higher precision along with the ResNet+LSTM model with data augmentation. However, the recall score is slightly different as the ResNet structure with data augmentation achieved 0.62 for no reaches (NoR) and 0.52 for reaches (R) which indicates that the number of false negatives is lower for the no reaches (NoR) cases. In contrast, the ResNet+LSTM with data augmentation had a recall of 0.51 for no reaches and 0.60 for reaches, thus leading to a lower number of false negatives for the case of reaches (Table 2.3).

Table 2.3: Comparative Results of the Performance of the Network Structures Considered

Model	Avg. Training Accuracy [%]	Avg. Validation Accuracy [%]	Avg. Testing Accuracy [%]	Precision NoR / R	Recall NoR / R
ResNet	98.30	50.61	53.43	0.65 / 0.40	0.55 / 0.51
ResNet+DA	94.59	53.65	58.16	0.68 / 0.45	0.62 / 0.52
ResNet+LSTM	98.61	50.59	54.21	0.65 / 0.41	0.57 / 0.49
ResNet+LSTM+DA	94.31	54.53	54.42	0.68 / 0.42	0.51 / 0.60
O-LSTM	75.44	81.16	63.71	0.59 / 0.82	0.92 / 0.35
MLP	47.66	46.13	51.8	0.55 / 0.67	0.78 / 0.42
BabyNet (Ours)	44.45	38.93	66.27	0.57 / 0.66	0.72 / 0.49

Results demonstrate that the proposed *BabyNet* outperforms all structures in terms of average testing accuracy while using the second smallest number of parameters and the small dataset of 63 reaches and in spite of featuring lower average training/validation accuracy. The observed lower validation accuracy can be explained by the fact that more challenging reaches were included in the validation set which, nevertheless, did not impact the ability of *BabyNet* to learn pertinent features. The observed lower training accuracy can be associated with the reduced training dataset size, whereby training accuracy would keep improving with more data. Still, *BabyNet* can perform well (in terms of testing accuracy) despite sub-optimal training.

Table 2.4: Comparative of the Trainable & Total Parameters of the Network Structures Considered

Model	Parameters (Trainable)/(Total)
ResNet	4,468,739 / 23,514,179
ResNet+Data Augm.	4,468,739 / 23,514,179
ResNet+LSTM	9,186,819 / 28,232,259
ResNet+LSTM+Data Augm.	9,186,819 / 28,232,259
O-LSTM	117,460,994 / 117,460,994
MLP	144 / 144
BabyNet (Ours)	1,204 / 1,204

The MLP structure is the smallest one (also uses 63 reaches), but it has the worst average testing accuracy (see Table 2.3). Furthermore, both networks predicted the same number of frames incorrectly during the reach phase but the MLP predicted 20

frames incorrectly during the no reach action whereas *BabyNet* only predicted six frames incorrectly. In term of keyframes, *BabyNet* only had a delay of one frame while predicting the reach whilst the MLP had a delay of four frames. In contrast, the MLP was not able to learn the motion of the reaching action, and had difficulties to discern the transition from a reach to a no reach.

2.3.4 Discussion

Evaluation results demonstrated that our proposed *BabyNet* is small yet powerful, and can challenge the performance of significantly larger structures by achieving 66.27% average testing accuracy (the highest one) on our dataset (See Table 2.4. The family of Resnet-based structures, despite their solid performance during training and validation, were found to provide results with increased false positives rates. On the other hand, the O-LSTM (that has the second best average testing accuracy and comparable to *BabyNet*'s) could not balance between no reaches (NoR) and reaches (R)—recall rates of 0.92 and 0.35, respectively. Yet, it remains an approach worth of further investigation in future work due to the ability of optical flow images to better differentiate subtle motion patterns compared to RGB images (see Figure 2.8).

Compared to the MLP (which is of comparable size to *BabyNet*), our *BabyNet* performed much better (approximately 27% improved performance) with almost the same precision/recall scores. However, it provides onset and offset keyframes at precision of one frame while the MLP had a delay of 4 frames.

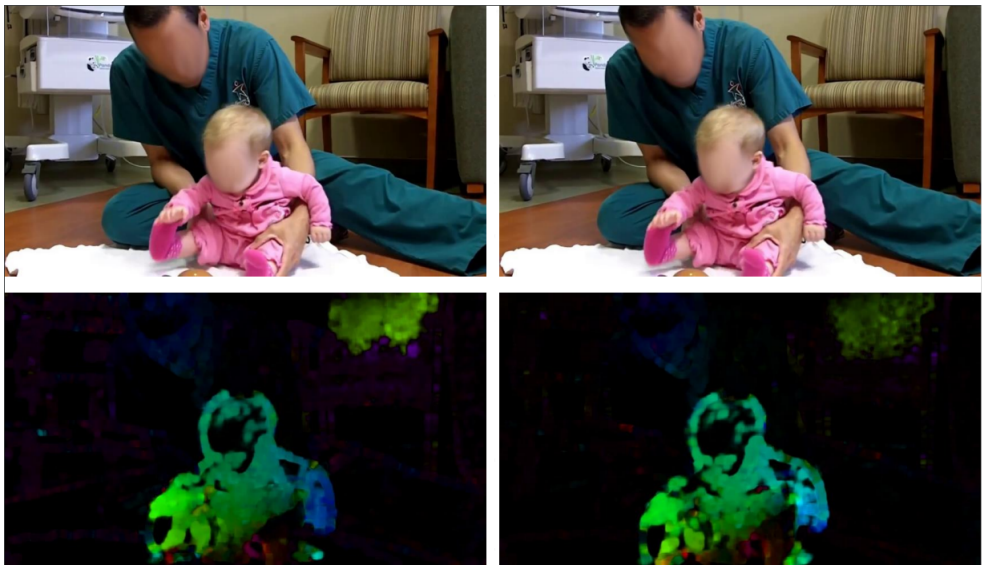

Figure 2.8: RGB (top panels) and equivalent optical flow images (bottom panels) spaced two frames apart (from left to right). Optical flow images can capture more clearly subtle changes, but at a (significant) pre-processing computational cost compared to plain RGB image inputs.

As this stage, *BabyNet* can be challenged by the lack of viewpoint variation in the dataset. Thus, extending the dataset introduced herein could generalize more challenging reaching action. The current framework lacks the capability to detect the reaching hand, an important information if the approach is to be integrated as an assessment tool or to facilitate the closure of the action-perception loop in forthcoming wearable devices. In this context, we introduce *E-BabyNet* as an extended framework building upon our initial reaching action recognition approach, *BabyNet*.

2.4 Approach II: *E-BabyNet*

2.4.1 Proposed Method

The proposed learning structure comprises two layers. The first layer is based on two LSTM models and aims at assimilating the correspondence between the hands' and

the object's bounding boxes using the distances and IOU. Each model assesses the reaching action for each hand by providing an output constituted of four scores: Onset (RN), Offset (RF), Reach (R), and No Reach (NoR) as described in Section 2.3.1. All frames that are between the Onset and the Offset are marked as Reach (R), whereas the frames before the Onset and after the offset are labeled as No Reach (NoR).

The two models are merged in the second layer through a Bidirectional LSTM (biLSTM). The latter processes the input in a forward and a backward direction. Thus, it considers past and future information critical when transitioning from a Reach (R) to a No Reach (NoR). The final output includes the four states RN, RF, R, and NoR along with Left (Lh) and Right (Rh) corresponding to the reaching hand. For example, a reach performed by a right hand will produce an output [Rh ↑, Lh ↓, RN ↑, RF ↓, R ↑, NoR ↓] during the onset phase and [Rh ↑, Lh ↓, RN ↓, RF ↑, R ↑, NoR ↓] during the offset phase, where the ↑ / ↓ arrows denote comparatively increased/decreased values.

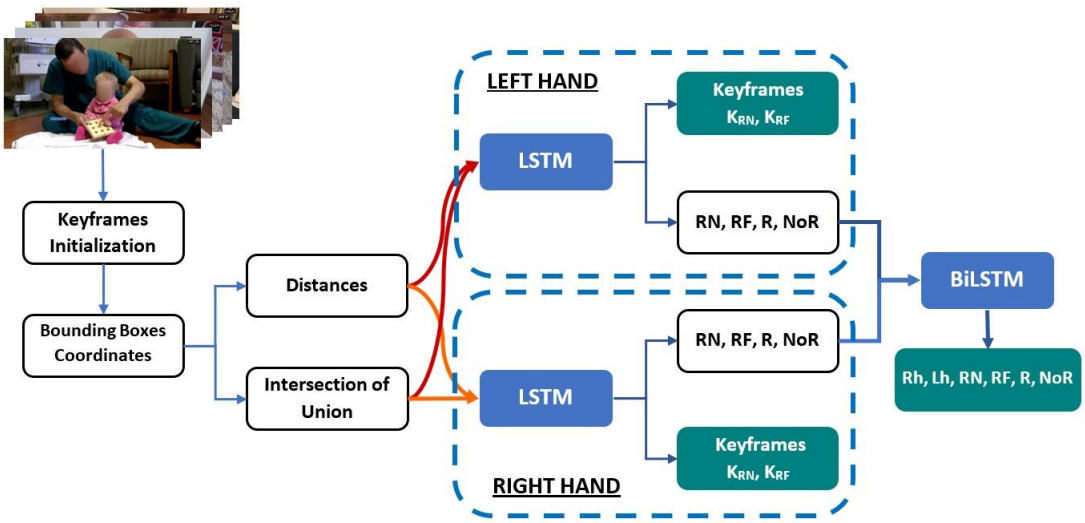

Figure 2.9: Flowchart of our approach *E-BabyNet*.

2.4.2 Implementation & Experiments

To evaluate our approach we considered a split of the dataset as follows: 70% training, 15% validation, and 15% testing using 266 reaches out of the total of 375. We also considered a final assessment using the remaining 109 reaches. All training, validation, and testing were performed on a workstation featuring an Intel Core i7 Processor (16 x 2.30GHz) and 16 GB DDR4 RAM, and an NVIDIA GeForce RTX 3050 GPU.

Based on previous results [37] that demonstrated the efficacy of comparatively smaller networks, the baselines in this work include only lightweight structures. These are

- Multi-Layer Perceptron (MLP),
- Gated Recurrent Unit (GRU),
- Bidirectional LSTM (BiLSTM), and
- *BabyNet*.

For each model, except for the *E-BabyNet*, the two inputs consist of the distances and the IOU of the detected hands' and objects' bounding boxes while the outputs are the scores (RN, RF, R, NoR) and the final keyframes K_{RN} and K_{RF}. As for the *E-BabyNet*, the inputs are processed separately for each hand (LH, RH) as shown in Figure 2.9 thereby also providing separate keyframes K_{RN} and K_{RF}. The final output constitutes of the scores (Rh, Lh, RN, RF, R, NoR).

2.4.3 Experimental Results

We present our experimental results during two phases: first, the implementation phase which includes a comparison of performance of the baselines structures with our approach *E-BabyNet* and a study of the effect of hyperparameters on the proposed approach, and next the evaluation phase of our method alone. In the implementation phase, we provide classification accuracy for training, validation and testing for each tested structure. In both phases, we evaluated the precision, recall and area under the receiver operating characteristic (ROC) curve (AUC) to gauge the performance of our approach. All networks were trained with learning rate 0.001 using Adam optimizer and cross-entropy loss with 30 hidden layers.

2.4.3.1 Implementation Phase

Comparison Against Baselines: Table 2.5 summarizes our findings for the first study that compares our method's performance against baselines. We observe that the MLP, GRU, *BabyNet* and BiLSTM baseline structures yielded the same performance with a significant difference in average accuracy for the MLP and a slightly lower recall score. With a testing accuracy of 60.3%, the recall scores were 0.61 for the MLP and nearly 0.70 for the rest of the structures. In contrast, our *E-BabyNet* offered high performance with a testing accuracy of 95.5% and recall and precision scores up to 0.96. The AUC confirms these findings as *E-BabyNet* has a score of 0.97.

Despite the similar performance in terms of precision and recall, predictions of keyframes were found to be quite distinct for the MLP compared to the other structures.

For all the correct reaching actions, the MLP had a difference within a range of five to eight frames in its keyframes. Moreover, it was unsuccessful at detecting comparatively shorter reaches (range of 3 to 7) compared to the three other baseline structures. *E-BabyNet* had a delay in a range of one to four frames but was able to yield fewer false positives.

Table 2.5: Performance of the *E-BabyNet* and Comparative Results Against Baselines

Model	Parameters	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
MLP	183	64.7	59.1	60.3	0.66	0.61	0.80
GRU	3153	72.6	72.3	85.6	0.68	0.70	0.82
BiLSTM	8103	76.3	77.7	85.5	0.71	0.69	0.81
BabyNet	3960	71.8	72.1	89.5	0.70	0.69	0.82
E-BabyNet	8249	95.4	97	95.5	0.96	0.97	0.97

Table 2.6: Hyperparameters Effect on the Performance of the *E-BabyNet*

Batch							
Batch	Hidden Layer	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
16	30	89.4	83.7	94.7	0.82	0.82	0.87
32	30	73.7	74.9	95.2	0.88	0.95	0.96
64	30	96.2	97.5	97.3	0.98	0.97	0.98
full	30	95.4	97	95.5	0.96	0.97	0.97
Hidden Layer							
Batch	Hidden Layer	Avg. Training Acc. [%]	Avg. Validation Acc. [%]	Avg. Testing Acc. [%]	Precision	Recall	AUC
full	15	61.5	61.0	57.6	0.82	0.82	0.87
full	20	72.6	72.3	85.6	0.68	0.70	0.82
full	30	95.4	97	95.5	0.96	0.97	0.97
full	100	99.9	100	100	1.0	1.0	1.0

Effect of Hyperparameters in E-BabyNet: Table 2.6 shows the results obtained during assessment of the effect that various key hyperparameters have on the performance of *E-BabyNet*. The key hyperparameters study includes varying hidden layer and batch values. During batch size evaluation, the structure maintained a good performance with lower accuracy and recall of 0.82. The full and 64 batch sizes performed best with high recall and precision scores; however, the full batch size needed over eight times less time to train, providing a better trade-off between performance and time execution. We observed that the hidden layer had an effect on the performance since the structure was overfitting with a hidden layer of 100 after 12 epochs. Based on these results, we infer that the structure with 30 hidden layers and full batch size is the most suitable for our application.

2.4.3.2 Evaluation Phase

We evaluated the performance of the *E-BabyNet* on a previously unseen set of 109 reaches. In this study, the first layer models for the left hand and the right hand were evaluated independently on the previously unseen dataset; the latter consists of 58.72% reaching actions lasting between 3 to 15 frames and 35.78% lasting between 16 to 30 frames.

Results are shown in Table 2.7. The precision and recall score were at 0.73 and 0.82, respectively. The reported keyframes had an average precision of three frames. The left hand model affected the performance of the overall structure with precision 0.69 and recall 0.89. We observed that 72% and 79% of reaching actions lasting between 3 to 7 and 8 to 15 frames were correctly recognized, respectively. In the case of challenging or confusing cases, the *E-BabyNet* is more prone to predict a false negative than a false positive. Lastly,

out of the six reaching actions with frames number greater than 31, four were correctly detected with a precision of four frames.

Table 2.7: Performance of the *E-BabyNet* During Evaluation Phase

Model	Precision	Recall	AUC
Left	0.69	0.89	0.86
Right	0.75	0.91	0.92
E-BabyNet	0.73	0.82	0.85

2.4.4 Discussion

Based on the obtained results we conclude that the *E-BabyNet* can outperform all baseline structures. Our structure also offered high performance in the final evaluation phase. The integration of the second layer, consisting of the biLSTM, is a key factor leading to high performance. Discontinued reaching actions were predicted regardless of the number of frames with partial information, e.g., the hands could be obstructed by the infant's body or another object in the scene during the reaching actions (Figure 2.10). Predictions of keyframes were also highly accurate, with some delays observed only in reaching actions that also included a final grasp of the object.

Results also suggest that the baseline structures provided low scores for precision which can lead to high false positives. Considering the purpose of the approach is to be implemented as a tool for the assessment of infants' actions, it is critical to prevent false positives to avoid inaccurate decision making mainly for pediatric applications. To this end, *E-BabyNet* is able to provide reliable action recognition, and when uncertainty is too high, it will render a false negative detection. Furthermore, while the baseline models underperformed in the case of short reaching actions, the *E-BabyNet* can recognize short

reaching actions as well as long ones. These findings demonstrate the efficiency of our approach and its ability to serve as an infant reaching action recognition tool.

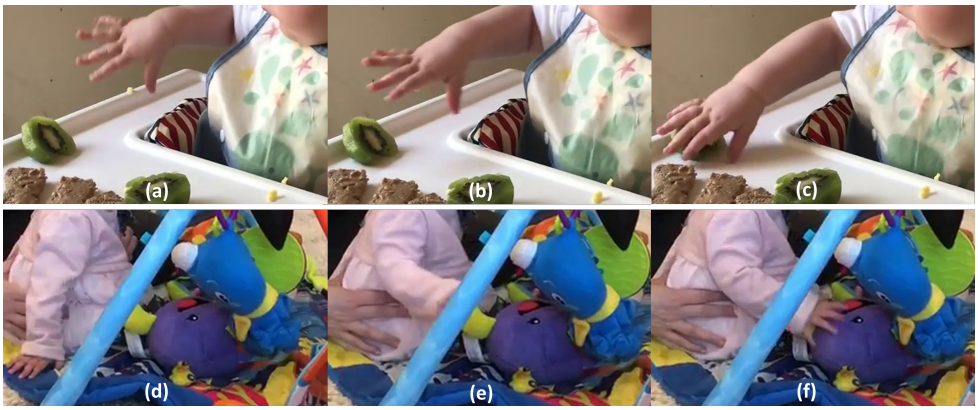

Figure 2.10: Examples of occlusion of the infant's hand during reaching actions. The top panels (a, b, and c) show the object obstructed by the infant's hand. The bottom panels (d, e, and f) show a hand obstructed by an object.

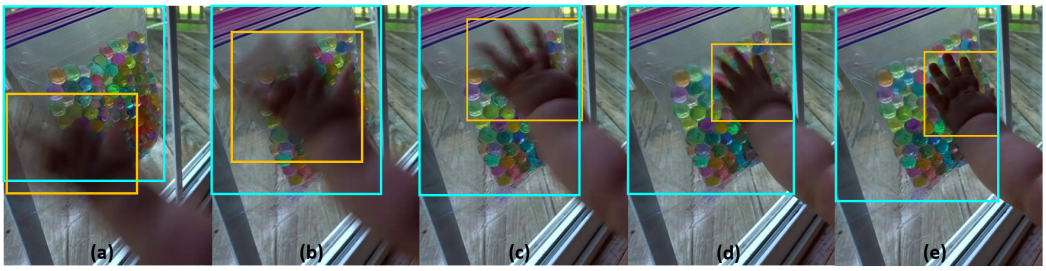

Figure 2.11: A reaching action with overlapping hand (orange) and object (cyan). (a) Onset phase, (b)-(d) during reaching action, and (e) Offset phase.

Despite these positive findings, the evaluation phase highlighted lingering challenges that can be further addressed. The front view reaching actions were the most challenging for the *E-BabyNet* as the bounding boxes of the hand and the object overlap at an early stage as shown in Figure 2.11. This effect was noticeable on the low score of the left model since left reaching actions were largely captured from the front and rear cameras view. In this case, the IOU is set at a high value during the onset phase and thus infers either a shorter reach or a no reach. As previously noted, the reaching action trajectories

are not straight and smooth until a particular stage of the infant's life, which can lead to cases where the distance between the hand and the object increases significantly throughout the reach action. This impacts the detection of the reaching if it occurs for more than three frames. Figure 2.12 illustrates two such sample trajectories. We also observed that several grasping actions were reported by the network as featuring a reduced number of frames compared to the correct one, owing to the fact that the infants' hand contacted the target object first prior to grasping it. This can be associated to the fact that our dataset includes training offsets with both touching and grasping which can lead to a few frames difference in the offset although the detection remains correct.

Figure 2.12: Examples of reaching action trajectories; top panels: 6 months old infant, and bottom panels: 7 months old infant. Panels (a) and (c) correspond to onset phases, while panels (b) and (d) show frames at the offset phase along with the complete trajectories followed by the hand.

2.5 Conclusions

In this chapter, we presented two novel light-weight networks called *BabyNet* and *E-BabyNet* aimed at infant reaching action recognition. *BabyNet* was tested using a newly

developed dataset containing 193 annotated reaches collected through 20 videos (available on the web via video-sharing platforms) from 21 different infants in this phase. The structure was found able to model short-range and long-range motion correlation of different key phases of a reaching action: its onset and offset. Overall, the performance provides a more accurate prediction compared to the one obtained from smaller-sized structures and is able to be competitive with larger structures as shown in Table 2.3. Additionally, *BabyNet* employs only a fragment of parameters to achieve these results as shown in Table 2.4.

Still, two key factors limiting the performance of our method include a lack of viewpoint variation in the dataset and the method's reliance on the detector network's effectiveness. Regarding the latter, missing detection of the hands and/or the object could compromise the whole reaching action recognition. In addition, *BabyNet* can indicate that a reaching is occurring without specifying the reaching hand which could be a key information for closing the action-perception loop in an upper extremity pediatric wearable robotic device. To address those limitations, we extended the structure to *E-BabyNet*, a two-layered learning structure able to assess separately right-handed and left-handed reaching action as well as to handle bimanual reaching which was previously difficult to detect. In addition, we integrated a Bidirectional LSTM (biLSTM) structure, which allows for better identification of the transition between the no-reaching and the reaching phases. Further, we augmented the reaching dataset to include more examples featuring bimanual reaching and viewpoint variations. The final dataset is constituted of 375 reaches performed by 40 different infants up to 12 months of age with bounding box annotations available. The annotations includes four classes: Infant, Left Hand, Right Hand, and Object.

Developing the infant reaching action recognition approach came with many challenges. A main challenge was the different field of view from the cameras across the videos. The camera angles at which the videos were recorded have a large influence on the perceived distance of the hand movement as it may quantify the infants range of motion. Another challenge is related to the variability of reaching actions that is seen in this population compared to adults. This can be attributed to the selected age where infants develop the reaching skill and the wide age range to include all levels of reachers (new vs. experienced).

This work also enables interesting future direction of research. First, the rich infant motion variability during development, for both typically-developing and neurodivergent subjects could be better harnessed by further extending of the reaching action dataset. To address challenging camera views, fusion with the infant's 2D skeleton data may improve the result of the first layer of the *E-BabyNet* structure, and thus increase the overall reaching action recognition efficiency. Furthermore, an online learning scheme could expand the scope of the structure and its capability to detect complex reaching actions, while a prediction model might aid in reducing the effect of hands' and objects' occlusions during reaching actions. In addition, the precision of the keyframes detection could be enhanced by an explicit definition of grasping and touching objects. Finally, we seek to implement and assess the performance of infant action recognition approach in closing action-perception loops of wearable robotic for upper extremity pediatric rehabilitation [84, 113, 114].

In the upcoming chapter, we will introduce another application of scene understanding focusing on precision agriculture.

Chapter 3

Leaf Detection and Pose Estimation in Support of Robotic Plant Phenotyping

Precision agriculture is an increasingly adopted farming practice that utilizes networks of ground and remote sensors to help improve use of agronomic inputs (e.g., water, fertilizers, pesticides) [180]. Current robots in precision agriculture predominantly focus on automated fruit and/or vegetable harvesting, along with utilizing remote sensing techniques for crop health monitoring. Comparatively less work has been performed with respect to collecting physical leaf samples in the field and retaining them for further analysis.

Though effective and critical, physical sampling techniques requires tedious and strenuous data collection procedures, henceforth limiting the number of samples gathered in practice. In turn, this undersampling can lead erroneous estimation of a crop's health

as only a handful of plants per are measured, and assessments are generalized to the entire plot [1]. Further, it is often the case that, depending on the crop, samples must be collected during a given time of the day to minimize measurement variations, which can further reduce spatio-temporal sampling frequency [1]. These challenges apply to leaf sampling too, which is the focus of this work. Enabling robotic leaf sampling for future use into phenotypic analysis processes can help improve measurement coverage and frequency while reducing human fatigue, risk of bodily injury, and labor costs. Besides remote sensing, an increasing number of works has begun addressing interactions with crops. Such works consider primarily robotic harvesting, including in row crops (such as corn and soybean) and in tree crops (such as citrus and avocado). For example, autonomous robots are deployed to pick sweet peppers, apples, citrus, and tomatoes by wrapping the fruit and twisting it off the stem with either a soft gripper [28, 68, 95], rigid gripper [32–34, 108, 109, 117, 155], or vacuum [10, 139, 179]. Some robots can pick strawberries, cucumbers, citrus, and sweet peppers by cutting the stem [7, 9, 65, 126, 156, 157].

Robotics research within the precision agriculture domain has largely focused on remote sensing via either ground or aerial robots (e.g., [82, 96, 127]). Leaf sampling is important in agriculture since remote sensing and monitoring typically provides field-level information without sufficient resolution to accurately diagnose problems. Compared to existing robotic leaf sampling methods [2, 5, 115, 120] and harvesting systems that cut the stem of a fruit/vegetable [7, 9, 65, 126, 156, 157], we are interested in performing clean cuts at leaves' stems and retaining leaves for further analysis. For this purpose, it is crucial to incorporate a visual perception component (to identify and localize a leaf) and an actuation

component (to move the end-effector toward the leaf, and then cut it). Collecting a leaf sample from a tree presents unique challenges in perception and actuation, different from robotic harvesting systems.

This chapter introduces a perception module for robotic means to aid in autonomously select, cut, and retrieve a leaf sample for future robotic leaf analysis. The developed perception algorithm is designed to identify and localize potential leaves and can be seamlessly incorporated into various perception-actuation framework. To support this claim, we combined our approach with two novel end-effector onto a six degree of freedom robotic arm to automate the leaf retrieval process. Results demonstrate the perception-actuation frameworks can successfully identify, localize, cut, and retrieve leaves.

The chapter's layout is as follows: in Section 3.1, we explore related works, while Section 3.2 outlines the development of the perception module using inputs from both 2D and 3D sources. Subsequently, Section 3.3 centers on the initial implementations and experiments related to leaf detection and pose estimation. The integration of our approach into a perception-actuation framework, accompanied by indoor and outdoor experiments and result discussions, is presented in Section 3.4. Lastly, Section 3.5 concludes the chapter.

3.1 Related Works

Visual perception can be used to monitor crop growth [133, 182], help prevent disease through early detection [6,77], assist with quality control [74,148], and help automate harvesting [18, 131]. The majority of systems are designed with a focus on tasks related to harvesting fruits and vegetables [7, 9, 11, 18, 92, 156, 157, 170], with only a handful of

exceptions [2, 5, 115, 120]. Mueller-Sim et al. demonstrated a robotic platform for rapid phenotyping that brought the laboratory to the field with the capability of manipulating leaves for in-situ measurement [2, 115]. Orol et al. developed a tele-operated aerial robot for cutting and collecting leaves from trees [120]. Ahlin et al. presented an algorithm for selecting and grasping tree leaves using a robotic arm [5]. The latter work demonstrates a high level of control using monoscopic depth analysis (MDA) and image based visual servoing, but focuses on grasping and pulling the leaf instead of cleanly cutting the stem of the leaf. Furthermore, these techniques have rarely been employed online on onboard computers as part of a robotic manipulation system to identify, localize and physically cut the leaf.

Simultaneously, non-destructive techniques methods are progressively being utilized, offering researchers and growers the means to acquire information regarding plant health, growth, and quality without causing damage to the crops. In this context, several works have been proposed for leaf segmentation, with a majority of them focusing on 2D image data due to its widespread availability and the early development of image segmentation methods. Kulikov [91] introduces an instance segmentation technique for leaf detection, employing images of solitary plants captured in the laboratory. His approach employs a two-step methodology, initially defining target embeddings, which are subsequently learned by a CNN. This enables a clustering approach during inference to recover individual instances. Weyler et al. [168] introduce a deep learning approach that predicts offset vectors directing to leaf and plant centers, followed by clustering to isolate individual leaves and plants. Guo et al. [60] directly predict leaf masks, bypassing the clustering post-processing

step. In contrast, our work centers on point clouds, as we aim to capture the 3D structure of plant leaves. Recent research has witnessed a surge in deep learning methods applied to process 3D data, utilizing techniques like voxel grids [29], multi-view image rendering [147], and convolutions on point clouds [153]. Similar to image-based methods, some of these approaches rely on offset vector estimation followed by point clustering [76, 161].

While the aforementioned works address segmentation within contexts such as autonomous driving and indoor settings, this dissertation focuses on agricultural field data, which presents distinct challenges, particularly stemming from scene deformability caused by meteorological factors. An inherent challenge in agricultural point cloud segmentation is the lack of publicly available datasets for instance segmentation. To address this, Schunk et al. [138] provide a dataset with tracked leaf instances over time, but it involves high-precision laboratory scanning, which is impractical for the field. In contrast, the approach in [107] operates directly on UAV-acquired point clouds in densely planted agricultural plots, encompassing a diverse range of leaf shapes influenced by real-world field conditions such as wind, heat, and lighting variations.

Still, these works are not culminating in tasks that encompass direct interaction with plants or crops, which is the main focus of this work. Enabling such tasks necessitates the implementation of pose estimation techniques. In order to successfully complete manipulation tasks such as physical leaves sampling, the 3D position of an object is insufficient. Thus, obtaining at least an estimate of the 6D pose (position and orientation) is critical. Traditional 6D pose estimation approaches usually perform local keypoint detection and feature matching, and then a RANSAC-based PnP algorithm on the established 3D-to-2D

correspondences to estimate the pose of an object [17, 111]. Still, they typically fail to perform with heavily occluded and poorly textured objects. On the other hand, learning-based methods use a deep neural network (DNN) to obtain the correspondences between 3D objects points and their 2D image projections [69, 70, 123]. Use of synthetic data generators [56, 183] can relieve in part the challenge of acquiring large labeled datasets; however, it requires realistic models that take into account the variations of the detected object e.g., shape, size, orientation or curvature which can be hard to develop.

3.2 Visual Perception for Leaf Detection & Pose Estimation

The development of the visual perception framework has progressed through various stages, ranging from the use of different cameras to the adoption of diverse approaches. This section details the different steps taken to develop our approach.

3.2.1 2D Object Detection and 3D Pose Estimation

In its early stages, the work centered on exploring the ZED mini camera alongside classical computer vision techniques leading to valuable insights.

The flow diagram of our initial visual perception algorithm is shown in Figure 3.1. The leaf detection approach is divided into two steps. First, both the left and right RGB and depth images are acquired from the ZED Mini stereo camera with a resolution of 1080 x 1920 and converted respectively to the HSV and YUV color spaces. The HSV conversion is set to highlight the green component in the RGB images as the YUV provides a better reading of the depth image through its Y component. We perform two-tiered segmentation, first

based on the HSV threshold to extract all tree leaves and then based on depth information to obtain leaves on the foreground. A dilation operation is applied on the obtained masked image to smooth the edges of the leaves.

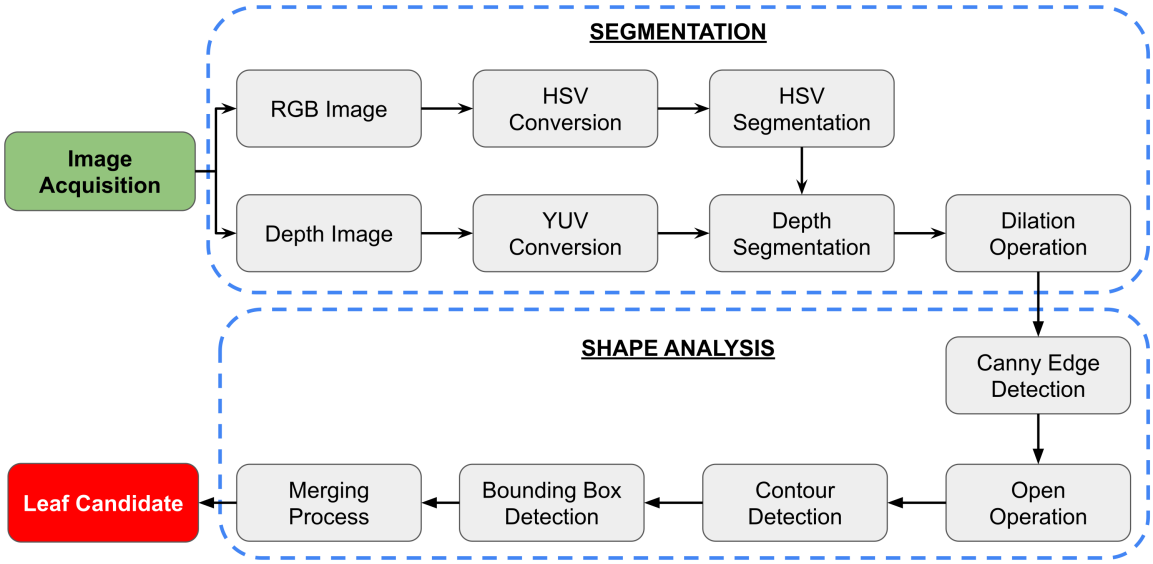

Figure 3.1: Flow diagram of the developed visual perception algorithm in initial development phase.

Next, the edges of the detected region are provided through a Canny edge detector [21] and processed with an opening morphological operation. From there, we perform a first classification process to extract the leaf contours information and retain only closed edges with high intensity and a maximum area for detecting bounding boxes. The output of the first classification is fed into a second classification, which is based on height/width ratio and orientation of the bounding boxes to provide the robot only with the accessible leaves. The parameters were selected through multiple trials and provide better performance than adaptive thresholding techniques such as Otsu's method [121].

Figure 3.2: Key instances and operations of the proposed 2D leaf detection algorithm. (a) Original RGB left image, (b) Threshold segmentation, (c) Depth segmentation, (d) Canny edge detection, (e) Contour detection, (f) Output bounding boxes and annotated keypoints.

Figure 3.2 depicts an example of a sequence of outputs from the various key different stages of our proposed detection algorithm. The final output includes the annotated stem, tip and leaf centroid, which are then used for leaf localization.

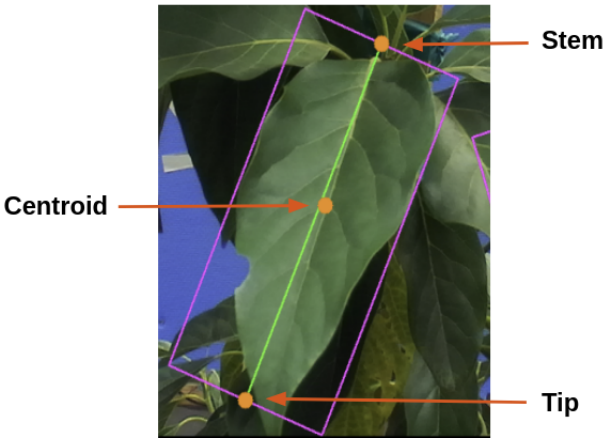

Figure 3.3: Estimated keypoints of a leaf candidate and detected bounding box.

To localize the leaf, we provide the estimated pixel coordinates from three parts of leaf candidates (stem, tip, and centroid keypoints) for left and right images within the left camera frame as shown in Figure 3.3. We estimate the 3D position directly from the 3D

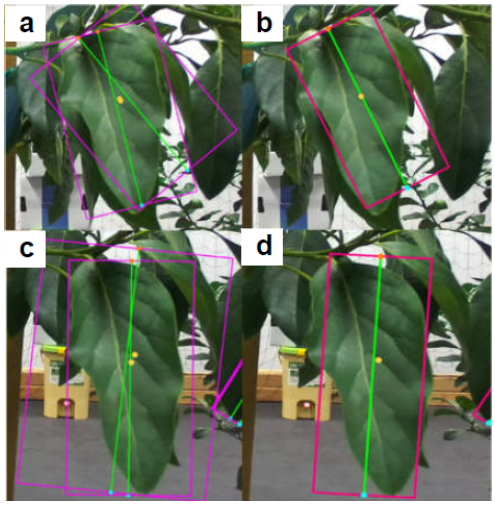

Figure 3.4: Panels (a) and (c): Examples of detecting bounding boxes twice. Panels ((b) and (d) Respective outputs to remove excess bounding boxed following our merging process.

point cloud provided by the ZED API. After the transformation from the pixel space to the spatial domain within the camera's reference frame, the leaf coordinates are transformed into the world coordinate frame. This reference frame transformation uses the known 6D-pose of the camera on the robot's end-effector to determine the position of the leaf within the world frame for retrieval.

In some trials bounding boxes were detected twice, as shown in Figure 3.4. These overlapping boxes can be explained by the reflection of the light on the leaves leading to the detection of double edges by the Canny detector. To rectify this, we applied a merging process that checks the distances between the detected centroids and based on the ratio between the related bounding boxes selects the most appropriate one.

At this stage, the successful picks were infrequent and retrieval often failed due to: 1) poor leaf centroid localization, 2) unreachable offset pose, or 3) leaf tip/end-effector misalignment. Consequently, one critical aspect to be addressed is the identification of the 6D-pose of the leaf which can improve the accuracy of the offset pose and remove the

necessity of assuming a fixed angle of attack to approach the leaf. An issue arising from the fixed angle of attack assumption is the possibility of leaves being pushed by the chamber's bottom side. In addition, leaf curling can lead to misalignment between the leaf's center vein "midrib") and the end-effector's direction of motion, which in turn may lead to sub-optimal stem cut or even pushing the leaf away from the cutting chamber.

3.2.2 3D Object Detection and 6D Pose Estimation

To perform any interaction, an accurate localization of the leaf is necessary. Though a 3D pose can be used, it is not sufficient to successfully accomplish the robotic task; thus, obtaining at least an estimate of the 6D pose is essential. For this purpose, we propose a leaf detection and localization algorithm using 3D point cloud and processed through the Open3D library. Our perception approach is outlined in Figure 3.5. The detection phase seeks to obtain the 3D bounding box of leaves candidates from point cloud captured from the depth camera. First, we remove outliers considered as noise resulting from sensor measurement inaccuracies and segment out the background at a specific distance threshold from the camera frame. Then, downsampling is applied to optimize the performance of the upcoming step. Next, we group the remaining point cloud segments into clusters using the Density Based Spatial Clustering of Applications with Noise (DBSCAN) approach [43]. It relies on two parameters, the minimum distance between two points to be considered as neighbors (eps) and the number of minimum points to form a cluster ($MinPoints$).

Each resulting cluster is considered a potential leaf and described by a 3D bounding box defined by center $C = [c_x, c_y, c_z]^T$, dimensions $D = [h, w, d]$, and orientation $R(\theta, \Phi, \alpha)$.

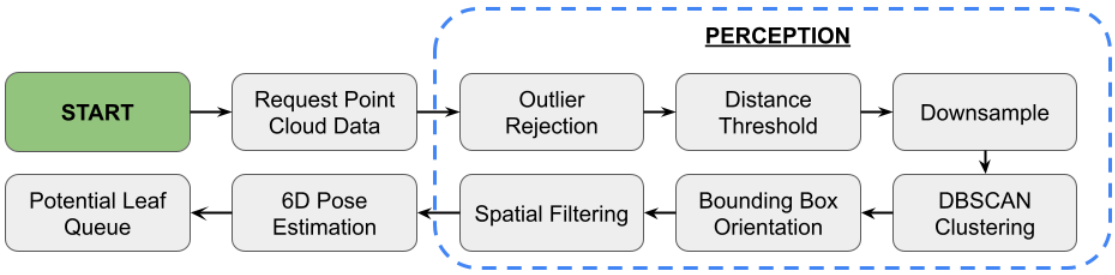

Figure 3.5: Flowchart of the proposed perception module. The point cloud data is processed to segment leaves and deposit leaf candidates into a queue.

Then, filtering is applied on the clusters using geometric features of the bounding box: number of points, volume, leaf ratio. Finally, the pose of the center of each bounding box is returned as the 6D pose of a potential leaf.

3.3 Preliminary Implementation & Experiments

In the next section, we detail the experiments carried out to validate our approach's effectiveness. This includes separate offline tests for both detection and localization, which were conducted.

3.3.1 Leaf Detection

For the detection step, ROSbags were collected both in indoor and outdoor settings. Indoors (lab with constant light conditions), we used the Kinova arm with the camera placed at different distances ($0.2 - 0.3$ m) from a potted tree. Outdoors (local orchard with varying light conditions), we collected data manually. We considered a wide range ($0.5 - 1.6$ m) of distances from trees; an example is shown in Figure 3.6.a. A total of

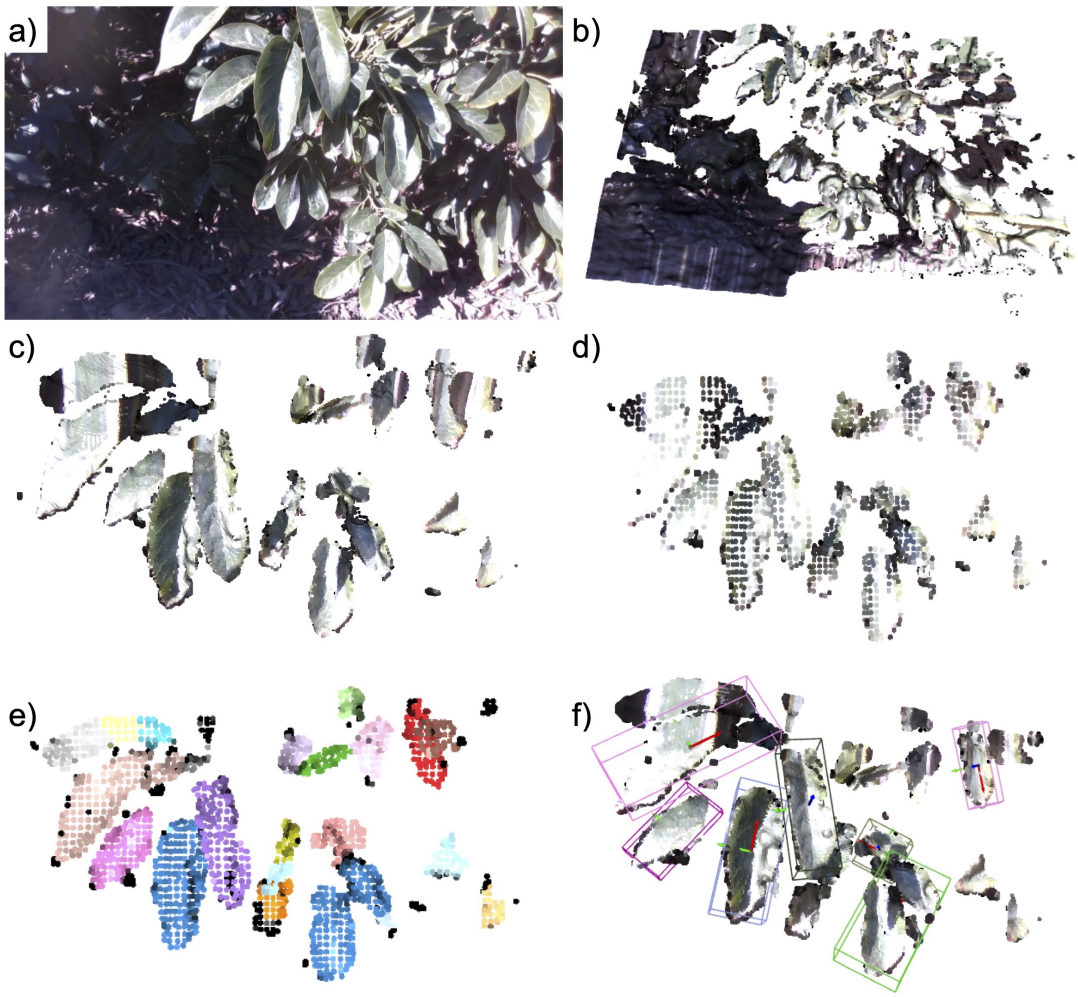

Figure 3.6: Key steps in our proposed leaf detection and localization process. The sample here corresponds to an outdoor point cloud: (a) corresponding RGB image of the tree, (b) raw point cloud, (c) distance filtered ROI, (d) downsampled point cloud, (e) segmented clusters, and (f) detected candidate leaves without 6D pose bounding boxes.

25 point clouds were collected (10 indoor and 15 outdoor). and tested offline with different combinations for *eps* and *MinPoints* parameters, to determine optimal values for later use.

Table 3.1 shows the outcome of our experiments on the 10 indoor point clouds and 15 outdoor point clouds. We attain an average of 80.0% of detection with a maximum of 90% for indoor dataset, and an average of 79.8% with a maximum 85% for outdoor. Further, we observed that the distance between the camera and the tree impacts the optimal values

for the point cloud processing. The greater the distance from the camera, the higher *eps* while *MinPoints* decreases.

Table 3.1: Leaf Point Cloud Detection

	Point Clouds	Total # Leaves	Average Detection	Percentage
Indoor	10	20	16	80.0%
Outdoor	15	99	79	79.8%

3.3.2 Leaf 6D Pose Estimation

To validate the localization phase, we compare several 6D poses obtained via our proposed approach against ground truth data obtained from a VICON motion capture camera system. Retroreflective markers were placed around the center of leaves, as shown in Figure 3.7, to estimate their pose.

Figure 3.7: We used motion capture to establish a ground truth for determining the leaf 6D pose. Markers were placed on a target leaf (left) with origin at the base of our 6-DOF robot (right). (A real avocado tree was used.)

Table 3.2 summarizes the results obtained for 12 random leaves positions. Our approach provides an estimation with mean error of 8.28 mm, 14.38 mm, and 15.54 mm along x-axis, y-axis, and z-axis, respectively,for avocado leaves of width ranging between $24 - 86$ mm and length ranging between $54 - 150$ mm. Based on the average leaf size (48×91 mm), estimation errors represent nearly 15% of the width and 17% of the length. We evaluated the orientation by calculating the Euclidean distance between the two provided values using the definition in [73]. We obtained a mean error of 5.3deg. We observe that the obtained 6D pose may drift from the physical center of the leaf mainly on the y-axis and z-axis due to human-induced error and the non-rigid nature of the leaf which impacts marker placement.

Table 3.2: Leaf 6D Pose Error

Error	Δx (mm)	Δy (mm)	Δz (mm)	Orientation (deg)
Mean	8.28	14.38	15.54	5.3
Std dev	7.46	5.46	6.69	15.5

The proposed approach provides an initial 6D pose along useful information of potential leaves using a processed 3D point cloud and obtained up to 80% of detection and a mean error less than 16 mm and 5.3 deg. Both detection and localization steps were performed without the need of collection of large data including 3D models, and training process. Furthermore, all tests were run using without any additional GPU acceleration.

3.4 Perception-Actuation Framework Integration

In order to better assess our pipeline for detecting and localizing leaves, we incorporated it into an actuation-perception framework designed for robotic phenotyping. Our experiments took place indoors using two distinct end-effectors, electrical and pneumatic, using an actual avocado tree [19, 39]. Following the validation of the outcomes, we proceeded to evaluate the pneumatic-based robotic system by conducting multiple real-world experiments in the Agricultural Experimental Station (AES) at the University of California, Riverside.

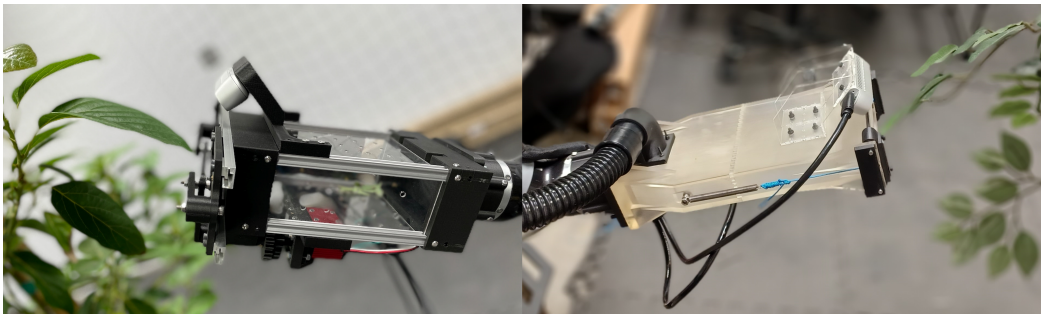

Figure 3.8: The electrical (left) and pneumatic (right) custom-built end-effectors.

3.4.1 Indoor Experiments

Two custom-built leaf-cutting end-effectors were developed and integrated independently. They were retrofitted on a mobile manipulation base platform (Kinova Gen-2 six degree of freedom [6-DOF] robot arm mounted on a Clearpath Robotics Husky wheeled robot). For perception, we utilize point cloud data from a depth camera (Intel RealSense D435i) for the leaf detection and localization algorithm developed (see Appendix A). The point cloud data is processed using Open3D [181] running on an Intel i7-10710U CPU, without any additional GPU acceleration.

We proceed with a detailed overview of the results achieved from the evaluation of these two distinct systems with a real potted avocado tree indoors. For every trial, the mobile manipulator and end-effector system was positioned at random poses near the base of the tree so that the end-effector was at distances ranging between $0.2 - 0.3$ m from the edge of the tree canopy. An experimental trial consisted of collecting a point cloud, storing the identified and localized potential leaves in a queue, and then sending the queued leaves to the arm for a retrieval attempt. Each trial concluded once the queue was depleted and the tree was repositioned for the next trial.

For each retrieval attempt, leaf candidates and viable leaves are determined. ***Leaf candidates*** are leaves that have a pose within the arm's workspace. ***Viable leaves*** are leaf candidates that have a retrieval path within the arm's workspace. For testing our point cloud detection, we are interested in monitoring both successful captures and successful cuts of the leaf. A ***successful capture*** occurs when the end-effector is placed around a viable leaf while a ***successful cut*** occurs when the enclosed leaf is removed from the tree. A ***clean cut*** occurs when the leaf is severed cleanly at the stem.

3.4.1.1 Electrical-based Robotic Leaf Retrieval System

Figure 3.9 highlights the integration of our contribution with the actuation module in a leaf-cutting system. The perception module processes point cloud data to segment leaves and deposit leaf candidates into a queue. Candidate leaves are then passed to the robot arm controller to actuate the end-effector using ROS velocity controller. These leaves act as the objective for the robotic arm to relocate and orient the end effector along a viable leaf, positioned at a distance away from the leaf's center known as the offset. This offset

corresponds to the leaf's length. Upon reaching this offset point, the arm proceeds to move linearly towards the leaf in order to grasp it. Upon successfully enclosing the leaf, the end effector performs a cut. Subsequently, the arm returns to its initial position. If the cutting process is unsuccessful, the arm controller requests the next leaf in the queue.

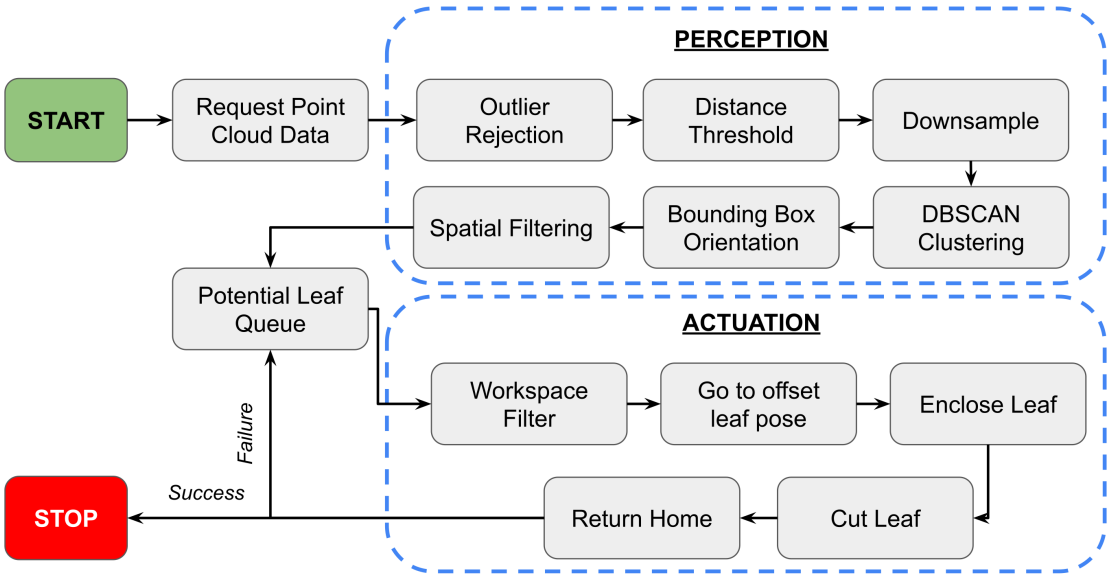

Figure 3.9: The overall integration of perception and actuation modules.

The developed stem-cutting end-effector utilizes two 4-bar linkages to actuate a set of sliding gates, one of which contains a razor blade to remove the leaf from the tree (Figure 3.10). The gates also help retain the leaf within the end-effector's chamber after removal from the tree. These 4-bar mechanisms are connected via a gear train to achieve synchronized motion. A low-cost, high-torque R/C servo (FEETECH FT5335M) drives the gear train while being amenable to position control. An Arduino Due microcontroller controls the servo motor and receives serial commands from a ROS control node. A breakout board connected to the Arduino contains a "safe/armed" switch along with LED indicators to reduce the risk of accidental injury. Finally, the RealSense D435i camera was positioned

adjacent to the opening of the end-effector, set at a 45° angle. This also serves to provide more depth when looking for potential candidates.

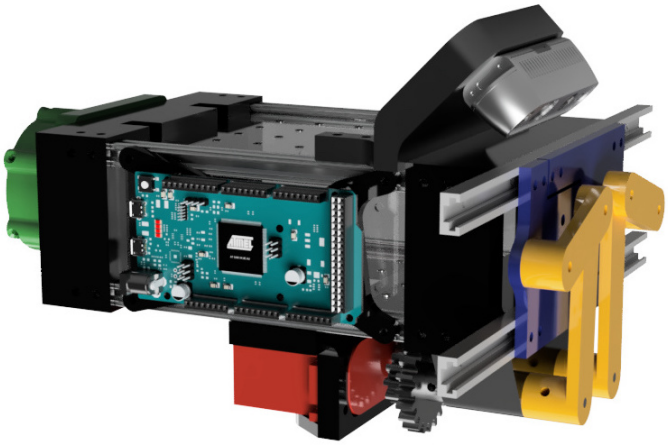

Figure 3.10: The electrical-based end-effector. The servo motor (red) actuates a double four-bar mechanism (yellow) that closes a set of gates (blue) with a razor blade to cut and capture a leaf. An Intel RealSense camera D435i is mounted on the top of the end-effector for perception. A microcontroller is mounted on the arm for controlling the motor. This end-effector can be mounted to a robotic arm using an adaptor plate (green). (Figure best viewed in color.)

Figure 3.11 outlines the overall process of a trial. Out of 46 trials, 63 potential leaves were detected by the point cloud. (Note that each point cloud in the trial could produce a variable amount of leaves, hence a higher number of potential leaves than trials.) After filtering the potential leaves to remove the leaves outside of the work space, 39 viable leaves remained. Out of these leaves, 27 were captured successfully (69.2%) while 21 of the 27 captured leaves were cut (77.8%).

Our system was able to remove a total of 21 leaves from the tree. However, not all leaves were clean cuts on the stem; four were classified as clean cuts for use in stem water potential analysis. The majority of the leaves were severed at the top of the leaf and not at the stem (Figure 3.12). Our system produced seven near-misses where the leaf

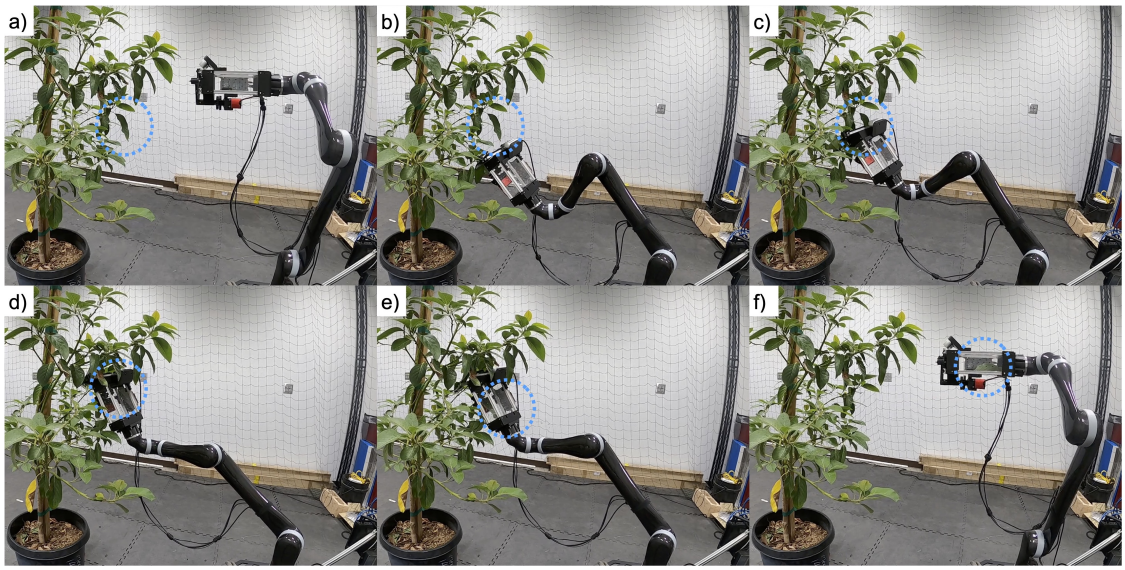

Figure 3.11: Overall leaf retrieval process. During the perception phase, (a) the point cloud is processed to determine a potential leaf. If a viable leaf is detected, (b) the arm will move to an offset position. (c) The arm will then perform a linear motion to capture the leaf. Once in position, (d) the arm will cut the leaf and (e) the leaf will fall into the enclosed chamber. (f) After completing the cut, the arm will return to the home position.

was cut within an average of 9.58 mm from the stem (std dev: 6.1 mm). The remaining 10 leaves were severed closer to the middle of the leaf, largely due to collisions with the branches. Similar branch interference also lead to four out of the six missed cuts from the captured leaf. These two problems could be solved through a refined end-effector design, more robust path planning to account for branches, and implementing visual servoing for continuous stem alignment as the end-effector approaches a viable leaf.

Table 3.3 summarizes retrieval results while Table 3.4 highlights the process times. The mean point cloud processing (perception) time was 5.6 sec and the mean cutting (actuation) time was 10.6 sec. The mean total retrieval time was 16.2 sec. The variation in processing time results from its dependence on multiple factors, notably the size of the point cloud, the spatial separation between the robot and the leaves, and the duration required for planning processing.

Table 3.3: Leaf Retrieval Numbers & Rates of the Electrical-based Retrieval System

Stage	Number	Rate
Potential Leaves	63	N/A
Candidate Leaves	51	81.0%
Viable Leaves	39	76.5%
Successful Captures	27	69.2%
Successful Cuts	21	77.8%
Clean Cuts	4	19.0%
Near Misses	7	30.0%

Table 3.4: Leaf Retrieval Performance Time (Seconds) of the Electrical-based Retrieval System

Metric	Perception Part	Actuation Part	Overall Retrieval
Min	0.5	4.6	6.1
Max	11.0	61.7	62.5
Mean	5.6	10.6	16.2
Median	7.7	8.1	15.3
Std dev	3.9	10.4	10.2

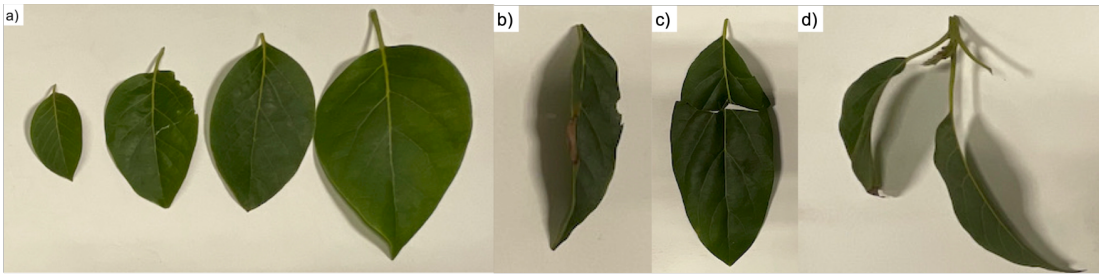

Figure 3.12: Sample leaves cut from our avocado tree during automated indoor tests. (a) The four leaves represent clean cuts suitable for stem water potential analysis. (b) The system also cut seven leaves that were classified as near-misses, which removed the leaf without the stem. (c) The remaining leaves were cut closer to the center, due to interference between the end-effector and the branches. (d) In two cases, collateral damage occurred when a second leaf was removed along with the target leaf. These instances were classified as a single successful cut, but not a clean cut since the two leaves would need to be separated for stem water potential analysis.

3.4.1.2 Pneumatic-based Robotic Leaf Retrieval System

To address the limitations of the previous system, we propose a framework that comprises three core components: actuation, perception, and communication. The new end-effector's design is shown in Figure 3.13. The dimensions of the opening and internal cavity of the end-effector were selected based on the average avocado leaf, allowing for one leaf to be captured while pushing other neighboring leaves aside.

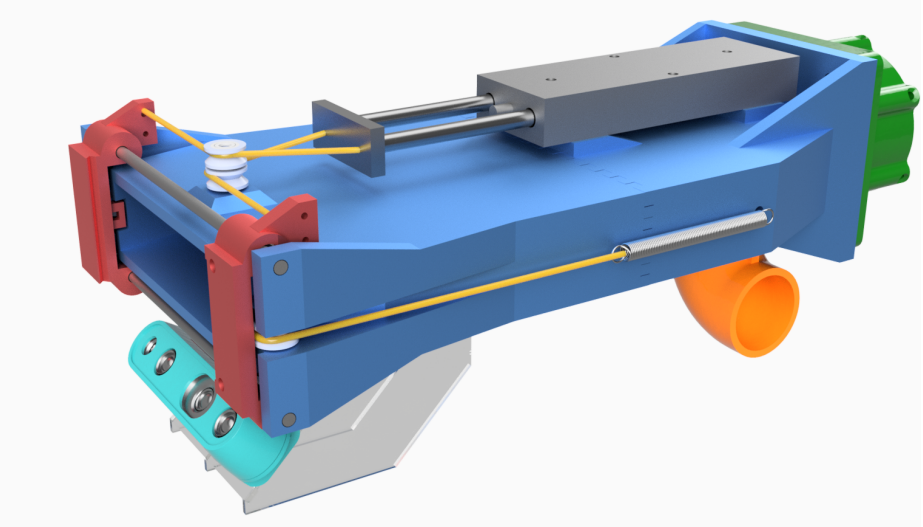

Figure 3.13: The pneumatic-based end-effector and its components. The end-effector body (blue) is mounted to the arm via the adaptor plate (green). The cutter (red) is mounted on linear rails and its actuated via the piston (grey) that is connected to tension strings (yellow) and pulleys (light grey). An Intel RealSense camera D435i (cyan) is mounted on the end-effector for perception. The suction tube connector (orange) is located at the back of the end-effector with direct access to the internal cavity. (Figure best viewed in color.)

To give a mechanical advantage to the leaf picking process, the opening of the end-effector can produce an inflow of air that guides and aligns the leaf toward the end-effector as it approaches the leaf, and so enhances the leaf picking algorithm's tolerance when dealing with misaligned leaf position and orientation. In addition, the RealSense

D435i camera was mounted near the opening of the end-effector at a 45° angle to allow for a clear view of leaves as they are retrieved.

The main body of the end-effector was manufactured with resin using stereolithography, which increased the permissible complexity of the design compared to other manufacturing methods and allowed us to take full advantage of precision additive manufacturing while using a rigid, lightweight, and non-porous material. This resulted in a stable end-effector that will not crack or break during field experiments.

Figure 3.14: Effect of the air suction mechanism on a leaf. The end-effector secures the tip of the leaf, allowing it to be captured even if the end-effector is misaligned or some misplacement is caused, e.g., by wind.

The cutter was designed to be compact and of low profile, with the ability to actuate with enough force and velocity to slice through the stem cleanly. This design is based on a pneumatic system. The latter consists of an air compressor, an air tank, and a solenoid. These are placed on the chassis of the robot while only guiding a single tube along the robotic arm, as shown in Figure 3.15.

To avoid exceeding the pressure rating of the pneumatic hardware, a pressure switch is connected in series with the load line of the compressor such that the switch opens

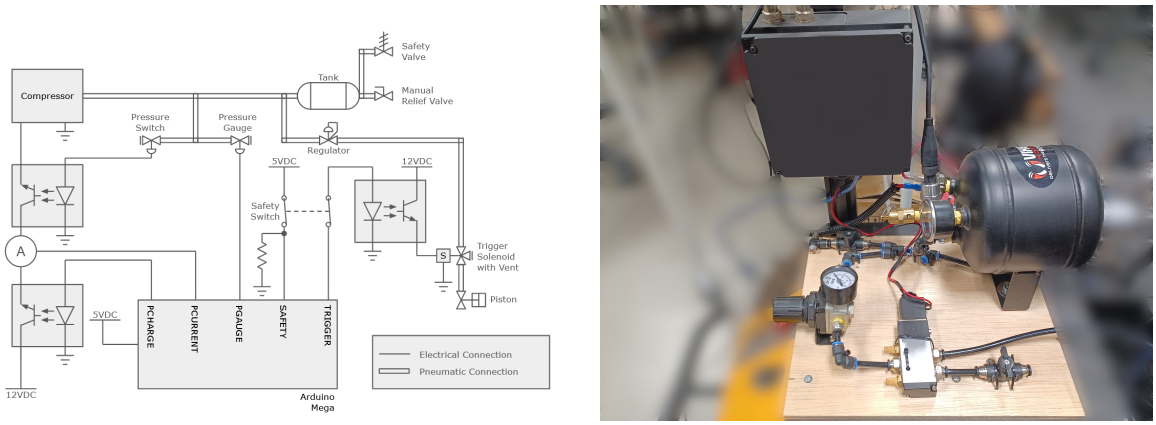

Figure 3.15: Electrical and Pneumatic Hardware. The detailed schematics of the pneumatic hardware (Left) and the hardware system mounted on the chassis of our mobile robot Husky (Clearpath).

above 120 psi. To actuate the cutting mechanism, a piston is linked to a pulley system; one of the sliders is equipped with a razor while the other one features a cavity to securely enclose the blade when the cutter closes. After a completed cut, the slider and piston position are reset using a set of springs connected to the pulley that retract the slider once the piston pressure is released.

The pneumatic subsystem is operated by a finite state machine through a microcontroller. The state machine may reject requests if it determines that the request would result in an invalid or dangerous state. Additionally, any state in this state machine may be immediately coerced into a safety "stop" state by a software signal, or by failing any of the safety checks that are routinely performed such as charge level.

A communication module facilitates the operation of the device and consists of a finite state machine that communicates with the various subsystems onboard. It is responsible for communicating with the perception module to initiate the point cloud capture, maintaining the queue of leaves that are to be harvested, and ensuring that the system is ready to perform cutting when the end-effector is in position.

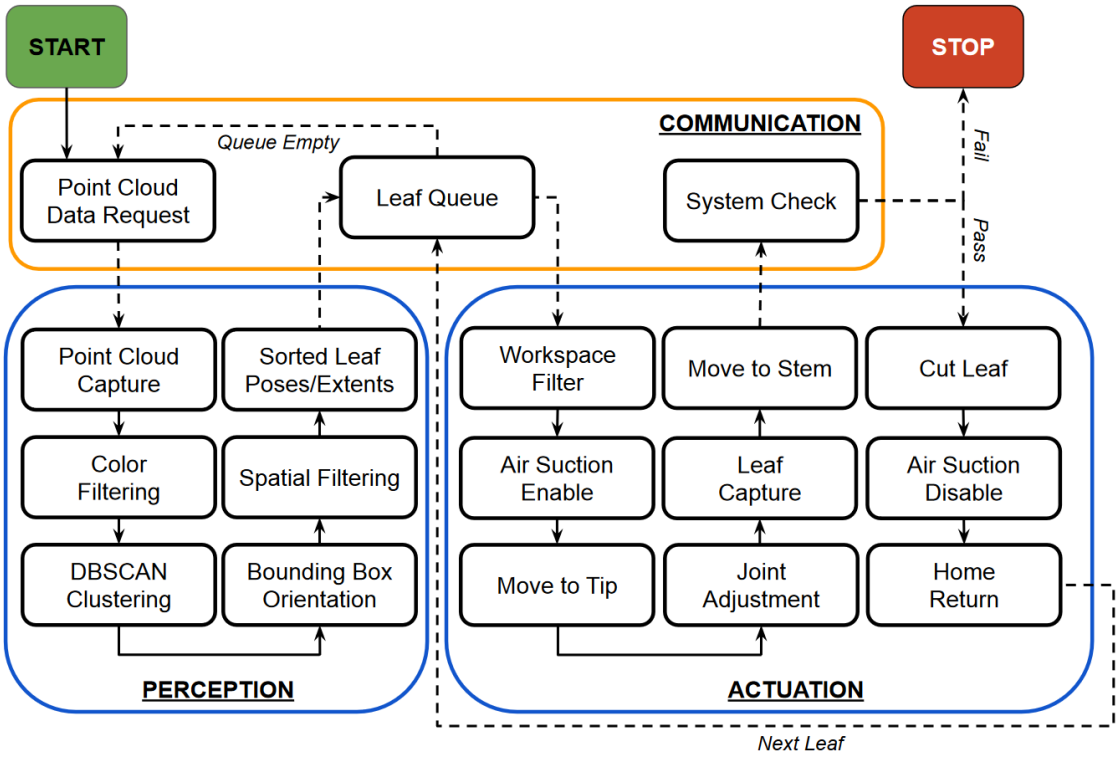

Figure 3.16: Our full system framework integrates actuation, communication, and perception module. The dashed arrows represent information flow.

When the system is initialized, a request is sent to the perception module to capture a point cloud and expects to receive back the poses and the dimensions of the leaves identified by the perception module. Then, the pose and dimension of the first leaf are sent as messages to the actuation system. These are interpreted into a leaf tip position and a stem position in the actuation module. Once the actuation system has operated the end-effector to move and capture the target leaf, the communication module confirms that the pneumatic subsystem is ready to actuate the razor. If this is the case, the motion sequence is activated. Once completed, the pose and dimensions of the next leaf in the queue are sent to the actuation module. If it is determined that the leaf queue has been

exhausted, a request for a new point cloud is sent. An overview of the complete system is provided in Figure 3.16.

With a few exceptions contributing to improved algorithmic efficiency, the perception approach employed herein adheres to the same paradigm as the one presented in Section 3.2.2. In sum, the point cloud is acquired with a background segmentation of 0.75 m from the camera frame and fully processed with no downsampling. We incorporated a color filtering operation aimed at excluding discolored leaves that may not qualify as suitable candidates for subsequent leaf analysis. Then, we apply the DBSCAN clustering algorithm and extract bounding boxes from the clustered point cloud and apply spatial filtering. Finally, we sort the potential leaves from closest to furthest based on their distance from the camera. As a result, more clusters are detected, and thus more leaf candidates are generated.

We conducted 42 trials with random position and orientation of the end-effector at a distance range of $0.2 - 0.4$ m. A total of 78 candidate leaves were obtained, 36 of which were viable. Of the 36 leaves, 27 were captured effectively (75%) whereas 25 out of 27 were removed (92.6%). With a 92% success rate, we acquired 23 clean cut stems with a minimum length of 5 mm as shown in Figure 3.17. Table 3.5 summarizes the retrieval results.

The processing times for leaf detection are 1.43 sec and 16.43 sec, respectively, and are explained by the process of the complete point clouds obtained, the size of which varies with the end-effector position. Furthermore, the motion planning has a minimum of 4.3 sec and a maximum of 25.84 sec, which corresponds to the variation in distance between the end-effector and the detected leaves.

Table 3.5: Leaf Retrieval Numbers & Rates of the Pneumatic-based Retrieval System

Stage	Number	Rate
Potential Leaves	132	N/A
Candidate Leaves	78	59.1%
Viable Leaves	36	46.1%
Successful Captures	27	75%
Successful Cuts	25	92.6%
Clean Cuts	23	92%
Near Misses	2	8%

Table 3.6: Leaf Retrieval Performance Time (Seconds) of pneumatic-based Retrieval System

Metric	Perception	Actuation	Full Retrieval
Min	1.43	25.33	37.00
Max	16.43	57.46	64.56
Mean	6.27	35.68	44.72
Median	6.18	34.83	44.20
Std dev	3.11	7.31	5.55

Figure 3.17: Leaves retrieved from the avocado tree during automated indoor tests.

3.4.2 Field Experiments

Based on the promising results obtained by the pneumatic-based framework, further field experiments were conducted at the Agricultural Experimental Station (AES) field at the University of California, Riverside.

First, 21 trials were performed with random location and orientation of the end-effector at a distance range of $0.2 - 0.5$ m. From a total of 90 candidate leaves, 21 proved to be viable. Out of the 21 leaves, 16 were efficiently enclosed (76.2%), while 10 were cut (62.6%). We acquired 6 neatly cut stems with a minimum stem length of 5 mm (see Table 3.7). The air suction system helped secure the capture of the leaf as we encountered some instances of wind during experiments. In terms of time performance, the mean time processing for the perception module is 9.27 sec and for the actuation module it is 37.34 sec, with an overall time processing of 46.61 sec as shown in Table 3.8.

The final step toward enabling autonomous leaf retrieval comprises the integration of waypoint navigation during deployment in real-world field experiments. To assess the robustness of our framework, consider a list of sentinel tree locations. This information is used to create desired waypoints that serve as the locations the robot should visit and attempt to sample leaves from the corresponding sentinel trees. Then, we integrate the leaf extraction steps outlined in Figure 3.16 for every encountered tree. When the leaf extraction process terminates for each sentinel tree, the robot proceeds to the next sampling area. This procedure continues until all designated sentinel trees have been visited, at which point the robot returns to its base.

Table 3.7: Leaf Retrieval Numbers & Rates

Stage	Number	Rate
Potential Leaves	193	N/A
Candidate Leaves	90	46.6%
Viable Leaves	21	23.33%
Successful Captures	16	76.2%
Successful Cuts	10	62.5%
Clean Cuts	6	60%
Near Misses	4	40%

Table 3.8: Leaf Retrieval Performance Time (Seconds)

Metric	Perception	Actuation	Full Retrieval
Min	4.29	27.57	38.70
Max	25.87	45.80	57.48
Mean	9.72	37.34	48.52
Median	9.10	36.33	48.93
Std dev	4.52	4.72	4.10

Field experiments were conducted in an avocado tree field at the Agricultural Experimental Station (AES; $33^\circ\ 58'\ 3.2592''\ N,\ 117^\circ 20'\ 7.0296'' W$) at the University of California, Riverside (see Figure 3.18. We use satellite imagery to construct an outline of the geometry of the field, including tree positions. Without loss of generality, we consider a case of sampling from three sentinel trees. Selected sentinel trees and the underlying Gaussian Processes (GP) reconstruction are shown in Figure 3.19. The computed path for the robot to follow as per the planning algorithm in [143] is also highlighted in the figure.

The satellite-based map is described in the World Geodetic System 1984 (WGS-84), but to be usable by the robot, it needs to be linked with the mobile robot's local map which is in turn used for robot navigation. We use the Universal Transverse Mercator (UTM) projection to express the satellite-based sampling points into desired waypoints in

the robot's local coordinate system. The mobile robot base (Clearpath Husky) can obtain odometry information from its wheel encoders, orientation, linear velocity and angular acceleration measurements from its embedded inertial measurement unit (IMU) module, and positioning data captured by its onboard Global Navigation Satellite System (GNSS) receiver. This information is used onboard in real-time by the $navsat\_transform\_node$[1] from the built-in ROS navigation stack, to broadcast the pose (i.e. position and orientation) of the mobile robot base in the UTM local coordinate system. The robot's pose is updated in the local frame while moving by using fused information from the three onboard sensory modules. The movement actions in the local frame are handled by the $move\_base$[2] ROS

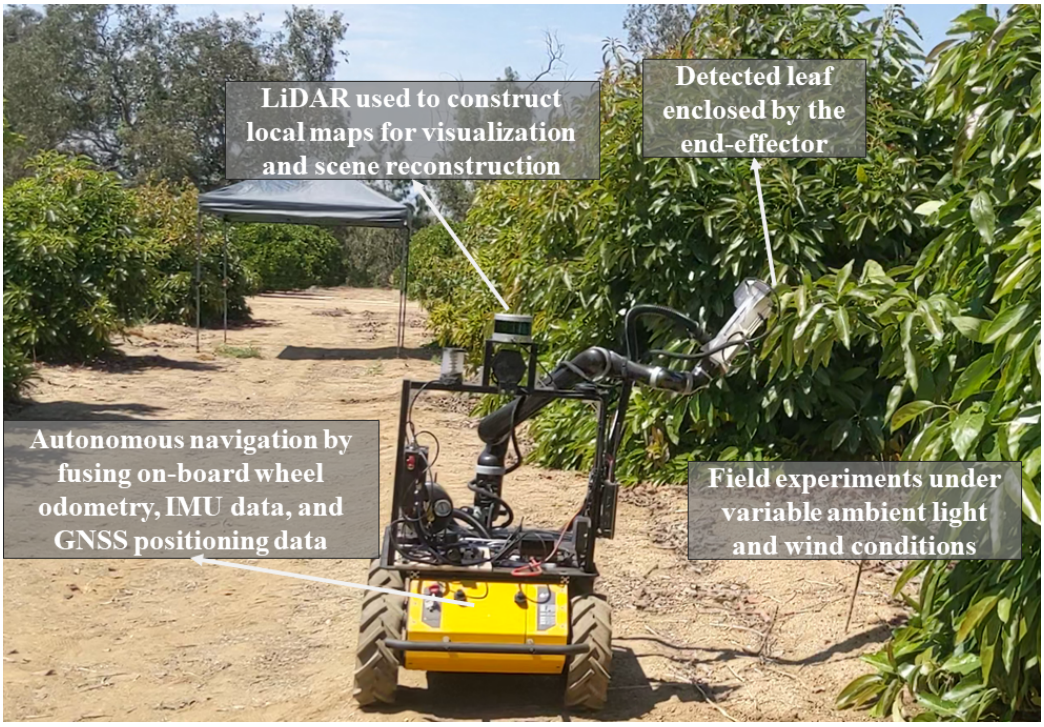

Figure 3.18: The agricultural robot used in this work for robotic assessment of our actuation-perception framework.

[1] `http://docs.ros.org/en/jade/api/robot_localization/html/navsat_transform_node.html`

[2] `http://wiki.ros.org/move_base`

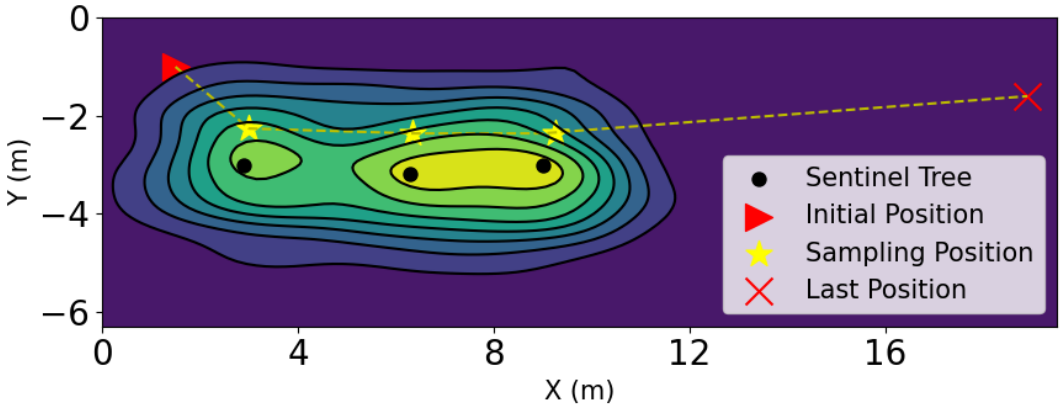

Figure 3.19: Gaussian Processes (GP) reconstruction of the avocado field with three sentinel trees considered during field experiments.

package, which generates velocity commands for the mobile platform in order to acquire the desired pose in the local frame. For additional safety, our developed system allows a human operator/supervisor to trigger when the robot switches between navigation and leaf-picking modes, as well as to skip a sampling location and move to the next one. The robot arm is rigidly affixed to the mobile base; coordinate transforms between the mobile base and each arm link frame as well the end-effector frame are all readily computed via closed-form forward kinematics expressions.

In our experiments, the mobile robot starts from a known position on the map. The robot arm is initialized turned toward the right-hand side of the mobile base, so that the camera mounted on the end-effector has an unobtrusive view to the tree canopies from the right-hand side of the robot as the latter moves forward.[3]

The first desired pose is transmitted to the platform, which in turn moves toward the target tree at the desired position using the generated trajectory. As the first goal pose is

[3] This configuration helps distribute the load to the mobile base as evenly as possible given other embedded parts, and minimizes occlusions to an embedded LiDAR sensor that is currently used to collect data during operation to create the visualizations shown in Figure 3.20).

reached, the mobile platform stops and the leaf retrieval process subsequently initiates. The perception module processes the collected point cloud and returns the center and dimensions of each candidate leaf with respect to the camera frame. The tip and stem positions are then estimated with respect to the end-effector's frame, and the manipulation planning procedure outlined in Figure 3.16 is executed. Figure 3.20(a) illustrates an instance of the mobile robot when sampling at the third desired location. Three candidates leaves have been successfully detected, and the process can then proceed to the actuation procedure. Figure 3.20(b) depicts the path followed by the robot in the field experiment until reaching the third sentinel tree. The complete field experiment can be viewed in the supplemental video at `https://youtu.be/xu4zrTe_S-U`.

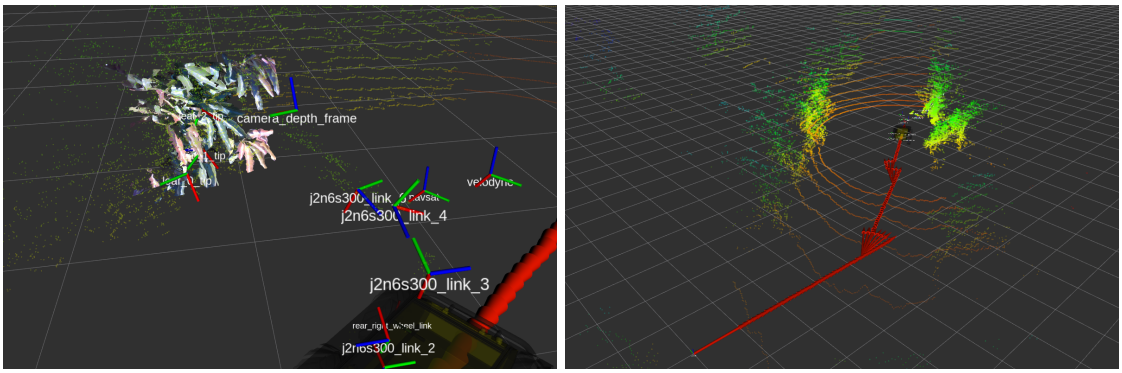

Figure 3.20: (a) Visualization (in ROS RViz) of the mobile robot at the third sentinel tree location. Each depicted coordinate system represents the corresponding state at the captured moment. Three leaf candidates, namely $leaf\_\{0,1,2\}\_tip$, have been detected. Given these candidates, the actuation module will decide to reach the closest one and attempt to cut and retain. (b) Visualization of the followed path in the avocado experimental field. The captured moment shows the robot in the third leaf sampling position, while at the leaf detection procedure. Red arrows illustrate the odometry poses along its path from the starting position.

3.4.3 Discussion of Collective Findings

The evaluation of the complete integrated framework for leaf-cutting demonstrates that the perception module can effectively adjust to various actuation modules, which may include robot arms and/or end-effectors.

The electrical-based actuation-perception system can capture 69.2% of viable leaves and cut 77.8% of those captured leaves. These results offered a promising initial step toward automated leaf sampling, nonetheless, they have revealed a few limitations that justify consideration. During the leaf-cutting experiments, we observed that some successful cuts were not accurate enough to be used for stem/leaf analysis (i.e. leaving long-enough stem length or damaging the leaf blade). One main issue was that the front face of the end-effector may push other interconnected leaves and/or branches away, hence the linear approach may not always suffice. In addition, high motion velocity can at times lead to misalignment with the enclosure and in turn to a failed capture and cut. Improving the alignment of the leaves can have the potential to significantly enhance the cutting process.

In order to tackle these limitations, we implemented the perception module with an optimized leaf-cutting end-effector that integrates minimal design with pneumatic suction to improve leaf cutting rates. In addition, we incorporated a color filtering technique to effectively exclude leaves that do not meet the criteria for being suitable candidates. The overall system was tested in indoor and outdoor settings.

In total, 43 leaves were successfully captured with an overall rate of 81.4% of successful cut with a mean time of 46.38 sec. 29 clean cuts were performed and only 6 near misses were encountered with an increase of 63.8% of clean cuts compared to the prior

framework. Furthermore, no missed cuts were encountered. These findings demonstrate the efficiency of the perception module in providing reliable position and orientation of the leaves but also dimensions that were critical to place the end-effector at the right position during the different phases of extraction. In few cases, leaves that are highly curved towards the midrib reduced the surface area of the detected leaf, which affects the location of the end-effector at the tip and may make the air suction less effective.

The proposed pneumatic-based end-effector design is able to capture leaves in dense areas of the tree with little interference from branches and other leaves. Furthermore, the air suction system assisted the capture process by securing the tip of the leaf within the end-effector regardless of misalignment due to partial detection, which proved to be practical during our field experiments as we observed the effect of wind on the position of the leaves. Despite its effectiveness, the overall system performance was impacted by various other factors. We observed instances where the planner fails at providing a feasible path for leaves in the workspace. This can be linked to the original position of the arm prior the process initiates, thus leading to limits angle joint during operation. Additionally, the planning time is in a range of 18.58 sec - 46.88 sec and motion time is in a range of 5.59 sec - 20.06 sec, showing the impact of the planning time on the overall actuation time as shown in Table 3.9.

Table 3.9: Actuation Performance Time (Seconds)

Metric	Planning	Motion	Overall Actuation
Min	18.58	5.59	27.57
Max	46.88	20.06	57.46
Mean	32.81	8.81	37.93
Median	31.88	8.23	37.71
Std dev	2.51	6.41	5.65

3.5 Conclusion

In this chapter, we introduced an adaptive perception module applied to different hardware designs to support robotic leaf phenotyping. We discussed the stages of development of the visual perception approach and highlighted the design of two distinct end-effectors, resulting in complete actuation-perception frameworks. The perception module demonstrated reliable detection and accurate 6D pose estimations leading to a solid performance from the presented frameworks. In controlled indoor settings, the pneumatic-based framework yields a 63.9% increase in cleans cuts compared to the electrical-based counterpart. Subsequently, field experiments were conducted to evaluate the performance under varying conditions such as lighting and wind. These experiments confirmed the effectiveness of our developed approach conducting experiments while the robot remains stationary and then guiding it through specific navigational positions. In the next chapter, we summarize the contributions of this dissertation and highlight potential future works.

Chapter 4

Summary & Directions for Future Works

To conclude this dissertation, we provide a concise overview of the main contributions from each chapter. Additionally, we highlight potential avenues for future research.

4.1 Infant action Recognition

Existing human action recognition algorithms are predominantly geared towards adult-oriented applications, rendering them less adaptable for pediatric scenarios. This trend is also observed in the predominant usage of datasets and methodologies that are tailored to adult movements, often failing to capture the distinctive characteristics of child behaviors and motions. Consequently, a limitation exists in the capacity of these algorithms to effectively analyze and comprehend the intricate actions of children. The presented work, in Chapter 2, is aimed at addressing and bridging this existing gap, by specifically considering the task of reaching which is an important developmental milestone.

First, we introduced *BabyNet*, a light-weight network structure to recognize infant reaching action from off-body stationary cameras. For this purpose, we developed an annotated dataset that includes diverse reaches performed while in a sitting posture by different infants in unconstrained environments (e.g., in home settings, etc.). Our approach uses the spatial and temporal connection of annotated bounding boxes to interpret onset and offset of reaching, and to detect a complete reaching action. We evaluated the efficiency of our proposed approach and compare its performance against other learning-based network structures in terms of capability of capturing temporal inter-dependencies and accuracy of detection of reaching onset and offset. Results indicate that our *BabyNet* can attain solid performance compared to other larger networks by achieving, and can hence serve as a light-weight data-driven framework for video-based infant reaching action recognition. However, the structure faced challenges due to the absence of diverse viewpoints in the dataset and the dependence on the detector network's performance. In cases where the hands or the object were not detected, the efficacy of the structure could be compromised. Moreover, the structure could not specify the hand performing the action.

To tackle certain aspects of these constraints, we developed *E-BabyNet* consisting of two main layers based on two LSTM and a Bidirectional LSTM (BiLSTM) model, respectively. The first layer provides a pre-evaluation of the reaching action for each hand by providing onset and offset keyframe based on *BabyNet*. Then, the BiLSTM model merges the previous outputs to deliver a final outcome of the reaching actions detection for each frame including the reaching hand. We evaluated our approach against four other lightweight structures using an extended and fully annotated dataset comprising 375 infant

reaching actions performed in sitting positions by different subjects. Results illustrate the effectiveness of our approach and ability to provide reliable reaching action detection and offer onset and offset keyframes with a precision of one frame. Moreover, the biLSTM layer handles the transition between reaching actions and reduced false detections.

This work also enables interesting future direction of research. The rich infant motion variability during development, for both typically-developing and neurodivergent subjects could be better harnessed by further extending of the reaching action dataset and include a broader range of neurodivergent subjects. To address challenging camera views, fusion with the infant's 2D skeleton data may improve the result of the first layer of the *E-BabyNet* structure, and thus increase the overall reaching action recognition efficiency. Furthermore, an online learning scheme could expand the scope of the structure and its capability to detect complex reaching actions, while a prediction model might aid in reducing the effect of hands' and objects' occlusions during reaching actions. Also, the precision of the keyframes detection could be enhanced by an explicit definition of grasping and touching.

These initiatives gear toward implementing and evaluating our methods in an on-going project to develop a wearable robotic device for pediatric upper extremity assistance, as described in [84, 113, 114]. Moreover, we harness the potential of both approaches as lightweight, data-driven machine-vision-assisted framework for the identification of early signs of neuromotor disorders that manifest themselves in delays with respect to typical developmental reaching milestones [129].

4.2 Object Detection and Pose Estimation for Robotic Plant Phenotyping

Contemporary robots in precision agriculture focus primarily on automated harvesting or remote sensing to monitor crop health. Comparatively less work has been performed with respect to collecting physical leaf samples in the field and retaining them for further analysis. While this technique offers benefits, the process of collecting, assessing, and interpreting measurements requires significant human labor and often leads to infrequent sampling. However, automated sampling can provide highly accurate and timely information. As a first step in such automated in-situ leaf collection, the process involve identifying and cutting a leaf from a tree. This retrieval process requires new methods for perception and actuation.

In chapter 3, we presented a technique for detecting and localizing candidate leaves using 3D point cloud from a depth camera. This technique was tested on both indoor and outdoor point clouds from avocado trees. We then employed two custom-built end-effectors, electrical- and pneumatic-based integrated onto a 6-DOF robotic arm to validate the proposed detection and localization technique by retrieving leaves from an avocado tree.

Indoor experimental testing with a real avocado tree showed that our perception module can enable our mobile manipulator and custom end-effector system to efficiently detect, localize, and cut leaves by successfully detecting an average of 80.0% of leaves indoors and 79.8% outdoors, and localizing them with less than 17% error along the leaf's length or width. The electrical-based actuation and perception framework for leaf-cutting achieved a capture rate of 69.2% for viable leaves, along with a cutting rate of 77.8% for

those captured leaves. While the end-effector demonstrates efficient leaf-cutting capabilities, its size becomes problematic when dealing with leaves clusters. This scenario highlighted the necessity for further design optimization to effectively tackle these challenges.

For this purpose, we introduced a pneumatic-based actuation module. Experimental indoor testing reveals that the overall framework is able to successfully captured 75% of viable leaves and successfully cut 92.6% of them. The new design enabled the capture of leaves in dense areas of the tree with little interference from branches and other leaves. Furthermore, the air suction system assisted the capture process by securing the tip of the leaf regardless of misalignment due to partial detection, which proved to be practical during our field experiments as we observed the effect of wind on the position of the leaves.

For future directions of research, we aim to extend the perception module to include visual feedback during capture; this could help handle various crop trees and thus various stem lengths, mainly in outdoor field testing. Furthermore, integrating an exploration strategy for detecting region of interest can lead to an optimized leaf sampling procedure. An automatic recovery procedure of the leaves is necessary as the air suction area was obstructed after three successful cut. In addition, incorporating a task planning strategy such as Next-Best-Action Planning (NBA-P) [79] or other online exploration techniques such as [78] would be highly advantageous for our robotic system. This would not only optimize the execution of additional tasks such as mapping and inspection but also enhance the overall efficiency of the system.

Bibliography

[1] Plant tissue sampling. `http://geisseler.ucdavis.edu/Guidelines/Plant_Tissue_Sampling.pdf`. Accessed: 2023-03-01.

[2] Justin Abel. In-field robotic leaf grasping and automated crop spectroscopy. Master's thesis, Carnegie Mellon University: Pittsburgh, PA, USA, 2018.

[3] Lars Adde, Annemette Brown, Christine Van Den Broeck, Kris DeCoen, Beate Horsberg Eriksen, Toril Fjørtoft, Daniel Groos, Espen Alexander F Ihlen, Siril Osland, Aurelie Pascal, et al. In-motion-app for remote general movement assessment: A multi-site observational study. *BMJ open*, 11(3), 2021.

[4] Karen E. Adolph and John M. Franchak. The development of motor behavior. *WIREs Cognitive Science*, 8(1-2), 2016.

[5] Konrad Ahlin, Benjamin Joffe, Ai-Ping Hu, Gary McMurray, and Nader Sadegh. Autonomous leaf picking using deep learning and visual-servoing. *IFAC-PapersOnLine*, 49:177–183, 2016.

[6] Tallha Akram, Syed Rameez Naqvi, Sajjad Ali Haider, and Muhammad Kamran. Towards real-time crops surveillance for disease classification: exploiting parallelism in computer vision. *Computers & Electrical Engineering*, 59:15–26, 2017.

[7] Christopher Aloisio, Ranjan Kumar Mishra, Chu-Yin Chang, and James English. Next generation image guided citrus fruit picker. In *IEEE International Conference on Technologies for Practical Robot Applications (TePRA)*, pages 37–41, 2012.

[8] Nabil Alshurafa, Wenyao Xu, Jason J Liu, Ming-Chun Huang, Bobak Mortazavi, Christian K Roberts, and Majid Sarrafzadeh. Designing a robust activity recognition framework for health and exergaming using wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1636–1646, 2014.

[9] Boaz Arad, Jos Balendonck, Ruud Barth, Ohad Ben-Shahar, Yael Edan, Thomas Hellström, Jochen Hemming, Polina Kurtser, Ola Ringdahl, Toon Tielen, and Bart van Tuijl. Development of a sweet pepper harvesting robot. *Journal of Field Robotics*, 37(6):1027–1039, 2020.

[10] Johan Baeten, Kevin Donné, Sven Boedrij, Wim Beckers, and Eric Claesen. Autonomous fruit picking machine: A robotic apple harvester. In *Field and Service Robotics*, pages 531–539. Springer, 2008.

[11] Guanjun Bao, Pengfei Yao, Shibo Cai, Shenshun Ying, and Qinghua Yang. Flexible pneumatic end-effector for agricultural robot: Design & experiment. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2175–2180, 2015.

[12] Neil E Berthier and Rachel Keen. Development of reaching in infancy. *Experimental Brain Research*, 169:507–518, 2006.

[13] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 850–865, 2016.

[14] A.N. Bhat and J.C. Galloway. Toy-oriented changes during early arm movements: Hand kinematics. *Infant Behavior and Development*, 29(3):358–372, 2006.

[15] Emily J. Blumenthal, Rain G. Bosworth, and Karen R. Dobkins. Fast development of global motion processing in human infants. *Journal of Vision*, 13(13):8–8, 11 2013.

[16] Tom Botterill, Scott Paulin, Richard Green, Samuel Williams, Jessica Lin, Valerie Saxton, Steven Mills, XiaoQi Chen, and Sam Corbett-Davies. A robot system for pruning grape vines. *Journal of Field Robotics*, 34(6):1100–1122, 2017.

[17] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3364–3372, 2016.

[18] Lingxin Bu, Guangrui Hu, Chengkun Chen, Adilet Sugirbay, and Jun Chen. Experimental and simulation analysis of optimum picking patterns for robotic apple harvesting. *Scientia Horticulturae*, 261:108937, 2020.

[19] Merrick Campbell, Amel Dechemi, and Konstantinos Karydis. An integrated actuation-perception framework for robotic leaf retrieval: Detection, localization, and cutting. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9210–9216, 2022.

[20] Merrick Campbell, Keran Ye, Elia Scudiero, and Konstantinos Karydis. A portable agricultural robot for continuous apparent soil electrical conductivity measurements to improve irrigation practices. In *International Conference on Automation Science and Engineering (CASE)*, pages 2228–2234. IEEE, 2021.

[21] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, 8(6):679–698, 1986.

[22] Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. *International Joint Conference on Artificial Intelligence (Special Track on AI for Good)*, 2022.

[23] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[24] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel Pierce, Daniel Bogen, Laura Prosser, Michelle J. Johnson, and Konrad P. Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020.

[25] Dimitrios Chatziparaschis, Elia Scudiero, and Konstantinos Karydis. Robot-assisted soil apparent electrical conductivity measurements in orchards, 2023.

[26] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[27] Steven W. Chen, Shreyas S. Shivakumar, Sandeep Dcunha, Jnaneshwar Das, Edidiong Okon, Chao Qu, Camillo J. Taylor, and Vijay Kumar. Counting apples and oranges with deep learning: A data-driven approach. *IEEE Robotics and Automation Letters*, 2(2):781–788, 2017.

[28] Girish Chowdhary, Mattia Gazzola, Girish Krishnan, Chinmay Soman, and Sarah Lovell. Soft robotics as an enabling technology for agroforestry practice and research. *Sustainability*, 11(23):6751, 2019.

[29] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR)*, pages 3075–3084, 2019.

[30] Jaired Collins, Joseph Warren, Mengxuan Ma, Rachel Proffitt, and Marjorie Skubic. Stroke patient daily activity observation system. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 844–848, 2017.

[31] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[32] Joseph R Davidson, Cameron J Hohimer, Changki Mo, and Manoj Karkee. Dual robot coordination for apple harvesting. In *ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2017.

[33] Joseph R Davidson, Abhisesh Silwal, Cameron J Hohimer, Manoj Karkee, Changki Mo, and Qin Zhang. Proof-of-concept of a robotic apple harvester. In *IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*, pages 634–639, 2016.

[34] Zhao De-An, Lv Jidong, Ji Wei, Zhang Ying, and Chen Yu. Design and control of an apple harvesting robot. *Biosystems Engineering*, 110(2):112–122, 2011.

[35] Roeland De Geest and Tinne Tuytelaars. Modeling temporal structure with lstm for online action detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1549–1557, 2018.

[36] Bappaditya Debnath, Mary O'brien, Motonori Yamaguchi, and Ardhendu Behera. A review of computer vision-based approaches for physical rehabilitation and assessment. *Multimedia Systems*, 28(1):209–239, 2022.

[37] Amel Dechemi, Vikarn Bhakri, Ipsita Sahin, Arjun Modi, Julya Mestas, Pamodya Peiris, Dannya Barrundia Enriquez, Elena Kokkoni, and Konstantinos Karydis. Babynet: A lightweight network for infant reaching action recognition in unconstrained environments to support future pediatric rehabilitation applications. In *International Conference on Robot & Human Interactive Communication*, pages 461–467, 2021.

[38] Amel Dechemi and Konstantinos Karydis. E-BabyNet: Enhanced Action Recognition of Infant Reaching in Unconstrained Environments. 2023. Under Review.

[39] Amel Dechemi, Olvera Hale Tomás, Christopher Eng, and Konstantinos Karydis. Design, Integration, and Field Evaluation of a New Pneumatic-based Robotic Leaf Cutting and Retrieving System. 2023. Under Review.

[40] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[41] Niki Efthymiou, Petros Koutras, Panagiotis Paraskevas Filntisis, Gerasimos Potamianos, and Petros Maragos. Multi-view fusion for action recognition in child-robot interaction. In *IEEE International Conference on Image Processing (ICIP)*, pages 455–459, 2018.

[42] Victor Emeli, Katelyn E. Fry, and Ayanna Howard. Towards Infant Kick Quality Detection to Support Physical Therapy and Early Detection of Cerebral Palsy: A Pilot Study. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1069–1074, 2020.

[43] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[44] Bjorg Fallang, Ola Didrik Saugstad, and Mijna Hadders-Algra. Goal directed reaching and postural control in supine position in healthy infants. *Behavioural brain research*, 115(1):9–18, 2000.

[45] Junming Fan, Pai Zheng, and Shufei Li. Vision-based holistic scene understanding towards proactive human–robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 75:102304, 2022.

[46] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano S. Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6016–6025, 2018.

[47] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image analysis*, pages 363–370. Springer, 2003.

[48] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10–10, 01 2007.

[49] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016.

[50] B. Fernando, E. Gavves, M. José Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5378–5387, 2015.

[51] Linda Fetters. Perspective on variability in the development of human action. *Physical therapy*, 90(12):1860–7, 12 2010.

[52] Longsheng Fu, Fangfang Gao, Jingzhu Wu, Rui Li, Manoj Karkee, and Qin Zhang. Application of consumer rgb-d cameras for fruit detection and localization in field: A critical review. *Computers and Electronics in Agriculture*, 177:105687, 2020.

[53] Hongwei Ge, Zehang Yan, Wenhao Yu, and Liang Sun. An attention mechanism based convolutional lstm network for video action recognition. *Multimedia Tools and Applications*, 78(14):20533–20556, 2019.

[54] Ahmed Ghali, Andrew S. Cunningham, and Tony P. Pridmore. Object and event recognition for stroke rehabilitation. In *Visual Communications and Image Processing*, volume 5150, pages 980 – 989. International Society for Optics and Photonics, 2003.

[55] Ross Girshick. Fast r-cnn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[56] Mario Valerio Giuffrida, Hanno Scharr, and Sotirios A. Tsaftaris. Arigan: Synthetic arabidopsis plants using generative adversarial network. *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2064–2071, 2017.

[57] Jun Goto, Takuya Kidokoro, Tomohiro Ogura, and Satoshi Suzuki. Activity recognition system for watching over infant children. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 473–477, 2013.

[58] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE/CVF International Conference on Computer Vision*, pages 5842–5850, 2017.

[59] Junxia Gu, Xiaoqing Ding, Shengjin Wang, and Youshou Wu. Action and gait recognition from recovered 3-d human joints. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1021–1033, 2010.

[60] Ruohao Guo, Liao Qu, Dantong Niu, Zhenbo Li, and Jun Yue. Leafmask: Towards greater accuracy on leaf segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1249–1258, 2021.

[61] Abdul Hafeez, Mohammed Aslam Husain, S.P. Singh, Anurag Chauhan, Mohd. Tauseef Khan, Navneet Kumar, Abhishek Chauhan, and S.K. Soni. Implementation of drone technology for farm monitoring & pesticide spraying: A review. *Information Processing in Agriculture*, 2022.

[62] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3154–3160, 2017.

[63] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[64] Jordan Hashemi, Mariano Tepper, Thiago Vallin Spina, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, Helen Egger, Geraldine Dawson, and Guillermo Sapiro. Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism Research and Treatment*, 2014, 2014.

[65] Shigehiko Hayashi, Kenta Shigematsu, Satoshi Yamamoto, Ken Kobayashi, Yasushi Kohno, Junzo Kamata, and Mitsutaka Kurita. Evaluation of a strawberry-harvesting robot in a field test. *Biosystems Engineering*, 105(2):160–171, 2010.

[66] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[67] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Raphael Weinberger, and A. Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and RGB-D data set. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 32–49, 2018.

[68] Cameron J Hohimer, Heng Wang, Santosh Bhusal, John Miller, Changki Mo, and Manoj Karkee. Design and field evaluation of a robotic apple harvesting system with a 3d-printed soft-robotic end-effector. *Transactions of the ASABE*, 62(2):405–414, 2019.

[69] Yinlin Hu, P. Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2020.

[70] Yinlin Hu, Joachim Hugonot, Pascal V. Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3380–3389, 2019.

[71] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4911–4920, 2023.

[72] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8981–8989, 2018.

[73] Du Q. Huynh. Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35:155–164, 2009.

[74] Gerhard Jahns, Henrik Møller Nielsen, and Wolfgang Paul. Measuring image analysis attributes and modelling fuzzy consumer aspects for tomato quality grading. *Computers and Electronics in Agriculture*, 31(1):17–29, 2001.

[75] Antoni Jaume-i Capó and Andreja Samčović. Vision-based interaction as an input of serious game for motor rehabilitation. In *IEEE Telecommunications Forum Telfor (TELFOR)*, pages 854–857, 2014.

[76] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2020.

[77] Peng Jiang, Yuehan Chen, Bin Liu, Dongjian He, and Chunquan Liang. Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks. *IEEE Access*, 7:59069–59080, 2019.

[78] Xinyue Kan, Hanzhe Teng, and Konstantinos Karydis. Online exploration and coverage planning in unknown obstacle-cluttered environments. *IEEE Robotics and Automation Letters*, 5(4):5969–5976, 2020.

[79] Xinyue Kan, Thomas C Thayer, Stefano Carpin, and Konstantinos Karydis. Task planning on stochastic aisle graphs for precision agriculture. *IEEE Robotics and Automation Letters*, 2021.

[80] Keren Kapach, Ehud Barnea, Rotem Mairon, Yael Edan, and Ohad Shahar. Computer vision for fruit harvesting robots—state of the art and challenges ahead. *International Journal of Computational Vision and Robotics*, 3:4–34, 2012.

[81] Manoj Karkee, Bikram Adhikari, Suraj Amatya, and Qin Zhang. Identification of pruning branches in tall spindle apple trees for automated pruning. *Computers and Electronics in Agriculture*, 103:127–135, 2014.

[82] Jeongeun Kim, Seungwon Kim, Chanyoung Ju, and Hyoung Il Son. Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications. *IEEE Access*, 7:105100–105115, 2019.

[83] Lauren Klein, Victor Ardulov, Yuhua Hu, Mohammad Soleymani, Alma Gharib, Barbara Thompson, Pat Levitt, and Maja J. Matarić. Incorporating Measures of Intermodal Coordination in Automated Analysis of Infant-Mother Interaction. In *International Conference on Multimodal Interaction*, page 287–295, 2020.

[84] Elena Kokkoni, Zhichao Liu, and Konstantinos Karydis. Development of a Soft Robotic Wearable Device to Assist Infant Reaching. *Journal of Engineering and Science in Medical Diagnostics and Therapy*, 3(2):021109, 2020.

[85] Elena Kokkoni, Effrosyni Mavroudi, Ashkan Zehfroosh, James C Galloway, Renè Vidal, Jeffrey Heinz, and Herbert G. Tanner. Gearing smart environments for pediatric motor rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 17(1), 2020.

[86] Jürgen Konczak, Maike Borutta, and Johannes Dichgans. The development of goal-directed reaching in infants II: Learning to produce task-adequate patterns of joint torque. *Experimental Brain Research*, 113:465–474, 1997.

[87] Jürgen Konczak, Maike Borutta, Helge Topka, and Johannes Dichgans. The development of goal-directed reaching in infants: hand trajectory formation and joint torque control. *Experimental Brain Research*, 106:156–168, 1995.

[88] Jürgen Konczak and Johannes Dichgans. The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental Brain Research*, 117:346–354, 1997.

[89] Jürgen Konczak and Johannes Dichgans. The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental Brain Research*, 117(2):346–354, 1997.

[90] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 2556–2563, 2011.

[91] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3843–3851, 2020.

[92] BongKi Lee, DongHwan Kam, ByeongRo Min, JiHo Hwa, and SeBu Oh. A vision servo system for automated harvest of sweet pepper in korean greenhouse environment. *Applied Sciences*, 9(12):2395, 2019.

[93] Do Kyeong Lee, Whitney G. Cole, Laura Golenia, and Karen E. Adolph. The cost of simplifying complex developmental phenomena: a new perspective on learning to walk. *Developmental Science*, 21(4):1–14, 2018.

[94] Joon Woong Lee, Mun Sang Kim, and In So Kweon. A kalman filter based visual tracking algorithm for an object moving in 3d. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 342–347 vol.1, 1995.

[95] Christopher Lehnert, Andrew English, Christopher McCool, Adam W Tow, and Tristan Perez. Autonomous sweet pepper harvesting for protected cropping systems. *IEEE Robotics and Automation Letters*, 2(2):872–879, 2017.

[96] Zhenbo Li, Ruohao Guo, Meng Li, Yaru Chen, and Guangyao Li. A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, 176:105672, 2020.

[97] Klaus Libertus, Jennifer Gibson, Nadia Z. Hidayatallah, Jane Hirtle, R. Alison Adcock, and Amy Needham. Size matters: How age and reaching experiences shape infants' preferences for different sized objects. *Infant Behavior and Development*, 36(2):189–198, 2013.

[98] Jiangjiang Liu, He Wang, R. Graham Cooks, and Zheng Ouyang. Leaf spray: Direct chemical analysis of plant material and living plants by mass spectrometry. *Analytical Chemistry*, 83(20):7608–7613, 2011.

[99] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[100] Shuai Liu, Dongye Liu, Gautam Srivastava, Dawid Połap, and Marcin Woźniak. Overview and methods of correlation filter algorithms in object tracking. *Complex & Intelligent Systems*, 7:1895–1917, 2021.

[101] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[102] MA Lobo, JC Galloway, and Jill Cathleen Heathcock. Characterization and intervention for upper extremity exploration & reaching behaviors in infancy. *Journal of Hand Therapy*, 28(2):114–125, 2015.

[103] Michele A Lobo and James C Galloway. The onset of reaching significantly impacts how infants explore both objects and their bodies. *Infant Behavior and Development*, 36(1):14–24, 2013.

[104] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[105] Yaqoob Majeed, Manoj Karkee, Qin Zhang, Longsheng Fu, and Matthew D. Whiting. Development and performance evaluation of a machine vision system and an integrated prototype for automated green shoot thinning in vineyards. *Journal of Field Robotics*, 38(6):898–916, 2021.

[106] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian Sminchisescu. 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2158–2167, 2018.

[107] Elias Marks, Matteo Sodano, Federico Magistri, Louis Wiesmann, Dhagash Desai, Rodrigo Marcuzzi, Jens Behley, and Cyrill Stachniss. High precision leaf instance segmentation for phenotyping in point clouds obtained under real field conditions. *IEEE Robotics and Automation Letters*, 8(8):4791–4798, 2023.

[108] Siddhartha S Mehta, William MacKunis, and Thomas F Burks. Robust visual servo control in the presence of fruit motion for robotic citrus harvesting. *Computers and Electronics in Agriculture*, 123:362–375, 2016.

[109] SS Mehta and TF Burks. Vision-based control of robotic manipulator for citrus harvesting. *Computers and Electronics in Agriculture*, 102:146–158, 2014.

[110] Ashish T Meshram, Anil V Vanalkar, Kavita B Kalambe, and Avinash M Badar. Pesticide spraying robot for precision agriculture: A categorical literature review and future trends. *Journal of Field Robotics*, 39(2):153–171, 2022.

[111] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 115–124, 2017.

[112] Longtao Mu, Gongpei Cui, Yadong Liu, Yongjie Cui, Longsheng Fu, and Yoshinori Gejima. Design and simulation of an integrated end-effector for picking kiwifruit by robot. *Information Processing in Agriculture*, 7(1):58–71, 2020.

[113] Caio Mucchiani, Zhichao Liu, Ipsita Sahin, Jared Dube, Linh Vu, Elena Kokkoni, and Konstantinos Karydis. Closed-loop position control of a pediatric soft robotic wearable device for upper extremity assistance. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1514–1519, 2022.

[114] Caio Mucchiani, Zhichao Liu, Ipsita Sahin, Elena Kokkoni, Konstantinos Karydis, et al. Robust generalized proportional integral control for trajectory tracking of soft actuators in a pediatric wearable assistive device. In *2032 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[115] Tim Mueller-Sim, Merritt Jenkins, Justin Abel, and George Kantor. The robotanist: A ground-based agricultural robot for high-throughput crop phenotyping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3639, 2017.

[116] Filipe Neves Dos Santos, Heber Miguel Placido Sobreira, Daniel Filipe Barros Campos, Raul Morais, Antonio Paulo Gomes Mendes Moreira, and Olga Maria Sousa Contente. Towards a reliable monitoring robot for mountain vineyards. In *IEEE International Conference on Autonomous Robot Systems and Competitions*, pages 37–43, 2015.

[117] Tien Thanh Nguyen, Erdal Kayacan, Josse De Baedemaeker, and Wouter Saeys. Task and motion planning for apple harvesting robot. *IFAC Proceedings Volumes*, 46(18):247–252, 2013.

[118] Tien Thanh Nguyen, Koenraad Vandevoorde, Niels Wouters, Erdal Kayacan, Josse G De Baerdemaeker, and Wouter Saeys. Detection of red and bicoloured apples on tree with an rgb-d camera. *Biosystems Engineering*, 146:33–44, 2016.

[119] Qiang Nie, Xin Wang, Jiangliu Wang, Manlin Wang, and Yunhui Liu. A child caring robot for the dangerous behavior detection based on the object recognition and human action recognition. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1921–1926, 2018.

[120] D. Orol, J. Das, L. Vacek, I. Orr, M. Paret, C. J. Taylor, and V. Kumar. An aerial phytobiopsy system: Design, evaluation, and lessons learned. In *International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 188–195, 2017.

[121] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[122] Carolina Pacheco, Effrosyni Mavroudi, Elena Kokkoni, Herbert G. Tanner, and René Vidal. A detection-based approach to multiview action classification in infants. In *International Conference on Pattern Recognition*, pages 6112–6119, 2021.

[123] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7667–7676, 2019.

[124] José Carlos Pulido, José Carlos González, Cristina Suárez-Mejías, Antonio Bandera, Pablo Bustos, and Fernando Fernández. Evaluating the child–robot interaction of the naotherapist platform in pediatric rehabilitation. *International Journal of Social Robotics*, 9(3):343–358, 2017.

[125] Qiu Quan, Tian Lanlan, Qiao Xiaojun, Jiang Kai, and Feng Qingchun. Selecting candidate regions of clustered tomato fruits under complex greenhouse scenes using rgb-d data. In *International Conference on Control, Automation and Robotics (ICCAR)*, pages 389–393, 2017.

[126] Redmond R Shamshiri, Cornelia Weltzien, Ibrahim A Hameed, Ian J Yule, Tony E Grift, Siva K Balasundram, Lenka Pitonakova, Desa Ahmad, and Girish Chowdhary. Research and development in agricultural robotics: A perspective of digital farming. *Chinese Society of Agricultural Engineering*, 2018.

[127] Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Thomas Lagkas, and Ioannis Moscholios. A compilation of uav applications for precision agriculture. *Computer Networks*, 172:107148, 2020.

[128] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[129] Nushka Remec, Judy Zhou, Joanne Shida-Tokeshi, Trevor A Pickering, Douglas L Vanderbilt, and Beth A Smith. Outcomes and hand use of reaching attempts: Comparison of infants at risk for developmental disability and infants with typical development. *Frontiers in Psychology*, 13:712252, 2022.

[130] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference of Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[131] Ali Roshanianfard and Noboru Noguchi. Pumpkin harvesting robotic end-effector. *Computers and Electronics in Agriculture*, 174:105503, 2020.

[132] Francisco Rovira-Más, Verónica Saiz-Rubio, and Andrés Cuenca-Cuenca. Augmented perception for agricultural robots navigation. *IEEE Sensors Journal*, 21(10):11712–11727, 2021.

[133] Pouria Sadeghi-Tehran, Kasra Sabermanesh, Nicolas Virlet, and Malcolm J Hawkesford. Automated method to determine two critical growth stages of wheat: heading and flowering. *Frontiers in Plant Science*, 8:252, 2017.

[134] Dimitrios Sakkos, Kevin D Mccay, Claire Marcroft, Nicholas D Embleton, Samiran Chattopadhyay, and Edmond SL Ho. Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy. *IEEE Access*, 9:94281–94292, 2021.

[135] Samuel Salazar-García, Isidro J.L. González-Durán, and Martha E. Ibarra-Estrada. Identification of the appropriate leaf sampling period for nutrient analysis in 'hass' avocado. *HortScience horts*, 50(1):130– 136, 2015.

[136] Sheila Schneiberg, Heidi Sveistrup, Bradford McFadyen, Patricia McKinley, and Mindy F. Levin. The development of coordination for reach-to-grasp movements in children. *Experimental Brain Research*, 146(2):142–154, 2002.

[137] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition (ICPR)*, pages 32–36, 2004.

[138] David Schunck, Federico Magistri, Radu Alexandru Rosu, André Cornelißen, Nived Chebrolu, Stefan Paulus, Jens Léon, Sven Behnke, Cyrill Stachniss, Heiner Kuhlmann, et al. Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *Plos one*, 16(8):e0256340, 2021.

[139] James Schupp, Tara Baugher, Edwin Winzeler, Melanie Schupp, and William Messner. Preliminary results with a vacuum assisted harvest system for apples. *Fruit Notes*, 76(4):1–5, 2011.

[140] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. On the estimation of children's poses. In *International Conference on Image Analysis and Processing (ICIAP)*, pages 410–421. Springer, 2017.

[141] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. On the estimation of children's poses. In *International Conference on Image Analysis and Processing*, volume 10485, pages 410–421, 2017.

[142] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[143] Azin Shamshirgaran and Stefano Carpin. Reconstructing a spatial field with an autonomous robot under a budget constraint. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8963–8970, 2022.

[144] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Conference on Neural Information Processing Systems (NeurIPS*, pages 568–576, 2014.

[145] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 1212.0402, 2012.

[146] Annette Stahl, Christian Schellewald, Øyvind Stavdahl, Ole Morten Aamo, Lars Adde, and Harald Kirkerod. An optical flow-based method to predict infantile cerebral palsy. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 20(4):605–614, 2012.

[147] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *International Conference on Computer Vision*, pages 945–953, 2015.

[148] Qinghua Su, Naoshi Kondo, Minzan Li, Hong Sun, Dimas Firmanda Al Riza, and Harshana Habaragamuwa. Potato quality grading based on machine vision and 3d shape analysis. *Computers and Electronics in Agriculture*, 152:261–268, 2018.

[149] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225, 2023.

[150] Satoshi Suzuki, Yasue Mitsukura, Hiroshi Igarashi, Harumi Kobayashi, and Fumio Harashima. Activity recognition for children using self-organizing map. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 653–658, 2012.

[151] Kalpit C. Thakkar and P. J. Narayanan. Part-based graph convolutional network for action recognition. In *British Machine Vision Conference (BMVC)*, 2018.

[152] Esther Thelen, Daniela Corbetta, and John P Spencer. Development of reaching during the first year: role of movement speed. *Journal of experimental psychology: human perception and performance*, 22(5):1059, 1996.

[153] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019.

[154] Antigoni Tsiami, Petros Koutras, Niki Efthymiou, Panagiotis Paraskevas Filntisis, Gerasimos Potamianos, and Petros Maragos. Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4585–4592, 2018.

[155] Naveen Kumar Uppalapati, B Walt, A Havens, Armeen Mahdian, Girish Chowdhary, and Girish Krishnan. A berry picking robot with a hybrid soft-rigid arm: Design and task space control. *Robotics: Science and Systems*, 2020.

[156] EJ Van Henten, BAJ van Van Tuijl, J Hemming, JG Kornet, J Bontsema, and EA Van Os. Field test of an autonomous cucumber picking robot. *Biosystems Engineering*, 86(3):305–313, 2003.

[157] EJ Van Henten, DA Van't Slot, CWJ Hol, and LG Van Willigenburg. Optimal manipulator design for a cucumber harvesting robot. *Computers and Electronics in Agriculture*, 65(2):247–257, 2009.

[158] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[159] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[160] Claes von Hofsten. Structuring of early reaching movements: a longitudinal study. *Journal of motor behavior*, 23(4):280–292, 1991.

[161] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2708–2717, 2022.

[162] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016.

[163] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[164] Tianhai Wang, Bin Chen, Zhenqian Zhang, Han Li, and Man Zhang. Applications of machine vision in agricultural robot navigation: A review. *Computers and Electronics in Agriculture*, 198:107085, 2022.

[165] Tianhai Wang, Yadong Liu, Minghui Wang, Qing Fan, Hongkun Tian, Xi Qiao, and Yanzhou Li. Applications of uas in crop biomass monitoring: A review. *Frontiers in Plant Science*, 12, 2021.

[166] David Webster and Ozkan Celik. Systematic review of kinect applications in elderly care and stroke rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 11(1):1–24, 2014.

[167] Tracy L. Westeyn, Gregory D. Abowd, Thad E. Starner, Jeremy M. Johnson, Peter W. Presti, and Kimberly A. Weaver. Monitoring children's developmental progress using augmented toys and activity recognition. *Personal and Ubiquitous Computing*, 16:169–191, 2011.

[168] Jan Weyler, Jan Quakernack, Philipp Lottes, Jens Behley, and Cyrill Stachniss. Joint plant and leaf instance segmentation on field-scale uav imagery. *IEEE Robotics and Automation Letters*, 7(2):3787–3794, 2022.

[169] Joshua L Williams and Daniela Corbetta. Assessing the impact of movement consequences on the development of early reaching in infancy. *Frontiers in Psychology*, 7:587, 2016.

[170] Ya Xiong, Yuanyue Ge, Lars Grimstad, and Pål J From. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. *Journal of Field Robotics*, 37(2):202–224, 2020.

[171] Ya Xiong, Cheng Peng, Lars Grimstad, Pål Johan From, and Volkan Isler. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Computers and Electronics in Agriculture*, 157:392–402, 2019.

[172] Hiroaki Yaguchi, Kotaro Nagahama, Takaomi Hasegawa, and Masayuki Inaba. Development of an autonomous tomato harvesting robot with rotational plucking gripper. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 652–657, 2016.

[173] Haibin Yan, Marcelo H Ang, and Aun Neow Poo. A survey on perception methods for human–robot interaction in social robots. *International Journal of Social Robotics*, 6:85–119, 2014.

[174] S. Yan, Yuanjun Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv*, 2018.

[175] Z. Yang, Y. Li, J. Yang, and J. Luo. Action recognition with spatio–temporal visual attention on skeleton image sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2405–2415, 2019.

[176] Yong Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[177] Feng-Ping Zhang, Frances Sussmilch, David S Nichols, Amanda A Cardoso, Timothy J Brodribb, and Scott A M McAdam. Leaves, not roots or floral tissue, are the main site of rapid, external pressure-induced ABA biosynthesis in angiosperms. *Journal of Experimental Botany*, 69(5):1261–1267, 2018.

[178] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H. S. Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *European Conference on Computer Vision (ECCV)*, pages 525—-542, 2020.

[179] Kaixiang Zhang, Kyle Lammers, Pengyu Chu, Zhaojian Li, and Renfu Lu. System design and control of an apple harvesting robot. *Mechatronics*, 79:102644, 2021.

[180] Naiqian Zhang, Maohua Wang, and Ning Wang. Precision agriculture—a worldwide overview. *Computers and Electronics in Agriculture*, 36(2):113–132, 2002.

[181] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv*, 2018.

[182] Yanjun Zhu, Zhiguo Cao, Hao Lu, Yanan Li, and Yang Xiao. In-field automatic observation of wheat heading stage using computer vision. *Biosystems Engineering*, 143:28–41, 2016.

[183] Yezi Zhu, Marc Aoun, Marcel Krijn, and Joaquin Vanschoren. Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants. In *British Machine Vision Conference (BMVC)*, page 324, 2018.

Appendix A

Camera Selection & Placement Evaluation

Several cameras were considered as the sensing modality for the proposed end-effector (Table A.1). Although the ZED and ZED2 have solid performance, they were excluded because of their wide baselines which do not fit our intended eye-on-hand configuration. We evaluated the performance of the three other cameras in different conditions including indoor and outdoor environments. The obtained results show that the Realsense (RS) D435i has the best performance, especially outdoors where it is able to provide a viable depth image at close ranges. Furthermore, we were able to obtain high-quality point clouds at depth ranges lower than those provided in manufacturer specifications (0.1 m). Sample images collected using the RS D435i are shown in Fig. A.1.

Two eye-on-hand configurations were considered, one looking straight ahead and one looking downward at a 45° angle. While the former case can lead to longer look-ahead distances, the latter one was ultimately selected. This configuration balances between pro-

Table A.1: Candidate Cameras Specifications

Camera	Baseline [mm]	Depth Range [m]	Field of View
ZED	120	0.3 – 25	90° x 60° x 100°
ZED2	120	0.3 – 20	110° x 70° x 120°
ZED mini	63	0.1 – 15	90° x 60° x 100°
RS D435i	50	0.2 – 3	87° x 58° x 95°
RS D455	95	0.4 – 6	87° x 58° x 95°

viding useful depth information about the tree (needed for obstacle avoidance and navigation around tree branches) and allowing for leaf detection and localization (needed for aligning the end-effector with the leaf to cut it).

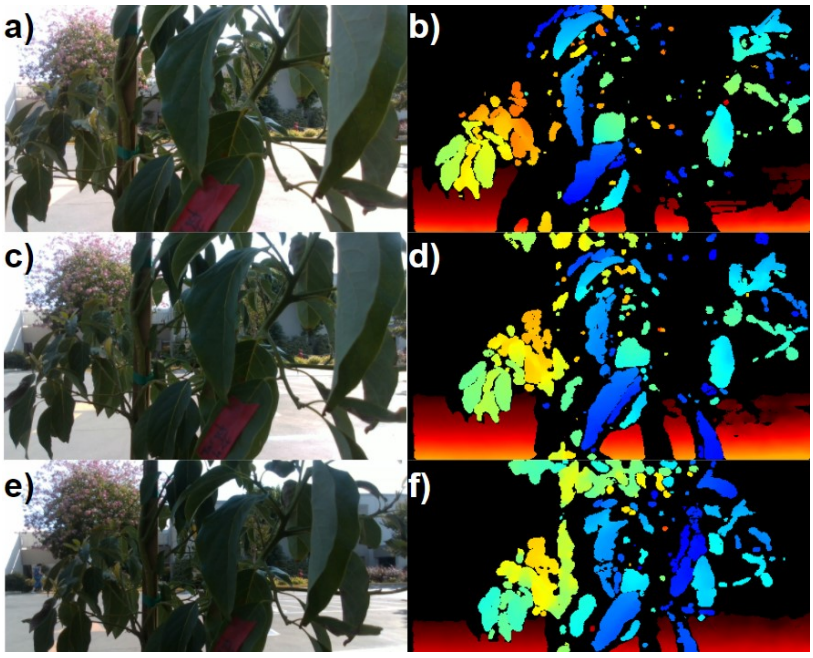

Figure A.1: Sample RGB and depth images collected from RS D435i in an outdoor environment at (a)–(b) 15 cm, (c)–(d) 20 cm, and (e)–(f) 25 cm.