University of Tennessee, Knoxville

TRACE: Tennessee Research and Creative Exchange

Masters Theses

Graduate School

5-2018

Peer Attention Modeling with Head Pose Trajectory Tracking Using Temporal Thermal Maps

Corey Michael Johnson
cjohn221@vols.utk.edu

Recommended Citation

Johnson, Corey Michael, "Peer Attention Modeling with Head Pose Trajectory Tracking Using Temporal Thermal Maps. " Master's Thesis, University of Tennessee, 2018.
https://trace.tennessee.edu/utk_gradthes/4980

To the Graduate Council:

I am submitting herewith a thesis written by Corey Michael Johnson entitled "Peer Attention Modeling with Head Pose Trajectory Tracking Using Temporal Thermal Maps." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Lynne E. Parker, Major Professor

We have read this thesis and recommend its acceptance:

Jens Gregor, Audris Mockus

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Peer Attention Modeling with Head Pose Trajectory Tracking Using Temporal Thermal Maps

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Corey Michael Johnson
May 2018

Acknowledgements

I would like to thank my advisor, Dr. Lynne E. Parker, for her excellent guidance and encouragement on this thesis. Also, my lab mates in the Distributed Intelligence Laboratory have been a tremendous resource for ideas and extra perspective. I would also like to thank all the mentors who invested time in teaching me over the years. Lastly, I've been immensely blessed with parents who have supported and nurtured my interests since childhood. Thank you all!

Abstract

Human head pose trajectories can represent a wealth of implicit information such as areas of attention, body language, potential future actions, and more. This signal is of high value for use in Human-Robot teams due to the implicit information encoded within it. Although team-based tasks require both explicit and implicit communication among peers, large team sizes, noisy environments, distance, and mission urgency can inhibit the frequency and quality of explicit communication. The goal for this thesis is to improve the capabilities of Human-Robot teams by making use of implicit communication. In support of this goal, the following hypotheses are investigated:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's Objects Of Interest (OOIs).

These hypotheses are investigated by experimentation with a new tool for peer attention modeling by Head Pose Trajectory Tracking using Temporal Thermal Maps (HPT4M). This system allows a robot Observing Agent (OA) to view a human teammate and temporally model their Regions Of Interest (ROIs) by generating a 3D thermal map based on the subject's head pose trajectory.

The findings in this work are that HPT4M can be used by an OA to contribute to a team search mission by implicitly discovering a human subject's OOI type, mapping the item's location within the searched space, and labeling the item's discovery state. Furthermore, this work discusses some of the discovered limitations of this technology and hurdles that must be overcome before implementing HPT4M in a reliable real-world system.

Finally, the techniques used in this work are provided as an open source Robot Operating System (ROS) node at github.com/HPT4M with the intent that it will aid other developers in the robotics community with improving Human-Robot teams. Furthermore, the proofs of principle and tools developed in this thesis are a foundational platform for deeper investigation in future research on improving Human-Robot teams via implicit communication techniques.

Table of Contents

List of Tables

List of Figures

1 Introduction

This chapter introduces and briefly covers the thesis motivation, problem statement, approach overview, challenges encountered, lessons learned, contributions, and overall thesis organization.

1.1 Motivation and Problem Statement

As robotics systems become cheaper and more capable over time, the ability to successfully pair humans and robots into cohesive teams could become a critical advantage for many new domains in the future. Effective robot team members will need the capability to implicitly discern current team goals and how they can best contribute without flooding the human members with requests for guidance. Also, the robot members will need to deal with different styles of human behavior, fit in naturally without causing unnecessary distractions, and adapt to dynamic situations. Many of these desired robot teamwork skills are currently hindered by the need for explicit communication. The challenge is for modern robots to reduce explicit communication and leverage implicit channels such as human body language. Figure 1 shows examples of modern Human-Robot teams.

Figure 1: Examples of Human-Robot Teams[1].

[1] Images sampled on 6/21/2017.
Left: Robonaut on the International Space Station NASA TV.
www.space.com/11127-nasa-space-station-robot-unpack-robonaut2.html .
Middle: Remotec hazardous duty robot.
www.lapatria.com/sucesos/la-policia-trabaja-de-la-mano-con-titus-165836 .
Right: Baxter robot learning a factory task.
www.rethinkrobotics.com/wp-content/uploads/2015/05/Baxter-Teach-Onexia.png .

The left image shows NASA's Robonaut [NASA-JSC, 2016] on the International Space Station (ISS). This system can work outside the ISS alongside astronauts and thus reduce the number of human crew members needed outside the safety of the space habitat. Although Robonaut can execute some maneuvers automatically, the system is typically commanded explicitly by a human. The center image shows a Remotec robot [Northrop Grumman International Inc., 2013] assisting with a bomb squad operation. This system allows technicians to view and operate on hazardous materials from a safe distance. Again, although it can perform certain operations automatically, most actions are explicitly commanded by a human. Lastly, the right image shows a factory robot system known as 'Baxter' [Rethink Robotics, 2017] learning a new task from a human. This system uses a computer screen with animated eyes to implicitly communicate to humans where the arms will move next.

With systems such as these in mind, it seems robots intended for human interaction could be more effective if the amount of explicit communication required for operations was reduced. These given motivations drive the following problem statement targeted for this thesis:

"*Develop a technology to improve Human-Robot teams by pushing the envelope of robot implicit communication on a real-world task such as a team search mission.*"

To constrain this research, this thesis studies this goal with the following robot system requirements:

- assist a human teammate with an object of interest (OOI) search mission,
- discover the human's OOI type implicitly,
- map the OOI locations,
- label the OOIs as being discovered or undiscovered by the human,
- map the candidate search frontiers,
- operate from a single robot unit perspective, and
- track the human head pose motions in 360°.

1.2 Approach Overview and Hypotheses

This thesis aims to prove the following hypotheses:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and

- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs.

This thesis investigates these hypotheses by implementing and conducting experiments with a new system for implicit peer attention modeling by Head Pose Trajectory Tracking using Temporal Thermal Maps (HPT4M). A new head pose tracking tool developed for this approach allows a robot Observing Agent (OA) to view a human teammate and temporally simulate their head gaze in 3D space. An RGB-D sensor is used to track the pose of a subject's helmet marker apparatus. This pose is then used to calculate and render a 3D temporal thermal map to represent saliency information about the human subject's attention zones. Subsequently, Regions Of Interest (ROIs) are extracted from this model and checked for OOIs. Finally, the OA maps the OOIs in the scene and labels whether or not the human teammate has discovered them.

1.3 Contributions and Lessons Learned

The contributions and lessons learned in this thesis are as follows. For scenes such as those set up in the experiment, HPT4M can be used to:

- Discover a subject's OOIs,
- Help find a subject's undiscovered OOIs, and
- Discover candidate frontier areas for search.

Also, the lessons learned about limitations and challenges of HPT4M are that it:

- Is limited by the RGB-D sensor physics (e.g., IR light reflecting on objects),
- Needs a priori knowledge of object shapes in order to render thermal maps correctly on unscanned backface surfaces, and
- Needs an accurate point cloud mesh algorithm that can render scene surfaces in 3D.

With these contributions and lessons learned in mind, this thesis claims that the HTP4M technology experimental results indicate proof of principle for implicitly aiding robotic agents in Human-Robot teams. It also shares valuable information with potential users regarding the limitations and inherent hurdles of this technology for real world applications. The broader impacts of robot systems making use of HPT4M could include:

- Reducing the amount of explicit commands that must be given by a human,
- Increasing the number of robots a human can manage at one time,

- Reducing the amount of time robot assets spend waiting for commands,
- Providing new application domains for Human-Robot Teams,
- Creating more intuitive, natural, and efficient interaction with humans, and
- Allowing human teammates more time to focus on higher level mission objectives.

Furthermore, the proofs of principle in this thesis are foundational work for deeper investigation in future related research. This research could include human trials with the team search mission, pairing the saliency ROI image stream with a general object classifier, and other applications for implicit temporal thermal based attention modeling in Human-Robot teams, such as human activity prediction.

1.4 Organization of the Thesis

This thesis first discusses related work on implicit communication in Human-Robot teams, human attention modeling and pose tracking in Chapter 2, then in Chapter 3 the HTP4M technical approaches for hardware and software are shown along with system validation testing. Chapter 4 covers the experiments for a Human-Robot team search mission, and temporal thermal mapping alternatives. Chapter 5 discusses the implementation challenges and limitations encountered. Finally, Chapter 6 summarizes the thesis conclusions.

2 Related Approaches and Sensor Systems

Implicit communication techniques applicable to Human-Robot team search missions are discussed first in Section 2.1. This information helps explain why implicit human attention modeling is chosen for investigation in this thesis and motivates the subsequent discussion of techniques for modeling and mapping human attention in Section 2.2. The success and widespread use of 2D eye tracking to model human attention directly inspires a new tool to be built for this thesis to investigate a 3D version of this technique based on head pose tracking. Consequently, a survey of image processing techniques and sensor systems for human head pose tracking is discussed in Section 2.3.

2.1 Implicit Communication in Human-Robot Teams

Human body language is a key signal for implicit communication in Human-Robot teams. In [Zhang et al., 2014], human body pose is used for implicitly recognizing activities such as running, walking, sitting, and falling. This signal can be acquired by extracting the human pose skeleton model with methods discussed in Section 2.3. Although this technique can allow a robot to recognize that a human teammate has started or stopped a search activity, other skill types are also needed for team search mission tasks. For example, as shown in [Breazeal et al., 2005], gaze behavior can be used to provide implicit social queues in teammate coordination. More information on this topic can be found in [Broz et al., 2015]. This technique helps facilitate efficient and natural communication between human and robot team members. Implicit gestures can also be used for implicitly directing a robot system to a new search location. However, this does not help a robot system to have intuition for discovering candidate search areas or current human OOI types on its own.

A critical implicit communication skill to pair with activity recognition and social queues for team search tasks is human attention modeling. Combining these different skills together can allow a robot system to: know that the search mission has started or stopped, be gesturally directed by a human teammate, and implicitly model the human's areas of attention for OOI mapping. This observation motivates the deeper investigation of implicit human attention modeling in this thesis. Thus, related approaches for implicit human attention modeling are discussed in Section 2.2.

2.2 Human Attention Modeling

2.2.1 Image Processing Techniques

A popular image processing technique for predicting human areas of attention is image saliency detection. This approach, as discussed in [Runxin et al., 2015], essentially maps areas in an image that are likely to stand out to a human observer. This is typically achieved by relying on image segmentation and color intensity in order to compute unique areas in an image that are likely to draw a human observer's attention. Although these techniques can help map where a human is likely to look, they do not actually map the human subject's real areas of attention. This characteristic causes the approach to be impractical for team search mission scenarios.

Other typical image processing techniques for modeling human attention are fundamentally derived from tracking eye gaze (see Section 2.2.2) or head pose (see Section 2.3.1). The capability to accurately acquire these signals from a rear perspective is not easily achievable for 2D image processing techniques due to absence of feature information on the back side of the human head or occlusion of the eyes. As shown in Chapter 3, this challenge is addressed in this work with a 3D image processing technique for localizing a standard marker target apparatus.

2.2.2 Mapping Attention Based on Eye Gaze

Eye gaze tracking is commonly used in the fields of psychology, neuroscience, marketing, and computer interfaces to model a human subject's interest in particular 2D area of a computer monitor screen. Duchowski's work [Duchowski, 2007] discusses eye gaze tracking applications such as these in depth. Thermal Mapping of Eye Gaze (TMEG) is a visualization technique to help highlight the subject's gaze locations over time. These thermal maps can be indicative of what the subject is thinking consciously or subconsciously, which can be critically useful information for exploration in these fields.

The company 'Package InSight' uses a TMEG system to indicating a shopper's visual interest in items for sale on the store shelves. An artificial rendering of their technology imposed over an image as projected in 3D space can be seen in Figure 2. This image alludes that marketing companies may potentially have interest in the novel capabilities of the HPT4M techniques developed in this thesis.

In Figure 3, the ‘Tobii Pro’ eye tracking system is being used to study the eye movements of the child observing a video of a person playing with two hand puppets. Studies such as these can help reveal details of human brain development and what age certain social capabilities and interests are acquired. Further details on eye tracking in developmental neuropsychology of children can be found in [Gredebäck et al., 2009]. The success of these systems to model human attention in 2D directly inspires the 3D technique developed in this thesis. As discussed in the next section, this design choice also directly motivates the need for a human head pose tracker.

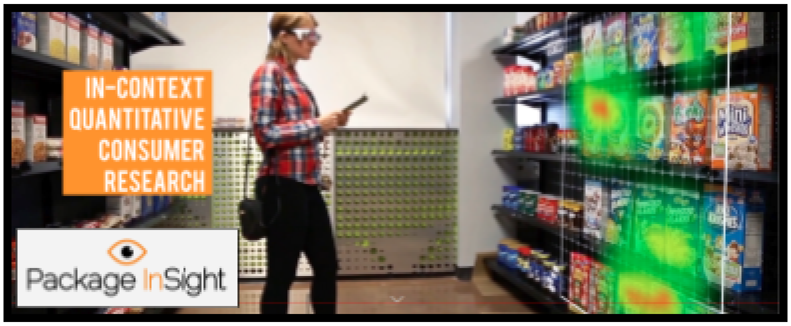


Figure 2: TMEG demonstration [PackageInSight, 2016].

Figure 3: Eye gaze tracking used in child development psychology [TobiiPro, 2017].

2.3 Human Head Pose Tracking Systems

An important prerequisite for computing a 3D attention map is an accurate head pose tracking system. The system requirement in Section 1.1 set the following requirements for the human head pose tracking system:

- Must operate from a single self-contained robot perspective, and
- Must track the human head pose motions in 360°.

As discussed in the next subsections, several systems exist for human head pose tracking, but none meet all of these requirements. Thus, a new tool for head pose trajectory tracking must be implemented in order to investigate the hypotheses in this thesis.

2.3.1 Computer Vision Techniques for Estimating Head Pose

A survey of head pose estimation techniques in computer vision can be found in [Murphy-Chutorian, and Trivedi, 2009]. This survey spans ninety papers and classifies the approaches into eight different taxonomies. These groups include methods such as: nonlinear regression, filtered image comparisons, kernelized subspaces, affine transforms, and manifold embedding. These techniques tend to focus on pose estimations from the frontal view because applications for frontal views are more common and rear head views are more challenging to track. Furthermore, these methods rely heavily on common information patterns of the human face and jaw line. The back side of human heads by comparison is bland and more variant. Figure 4 shows a typical head pose estimation technique based on face feature matching [Sung et al., 2008]. This strategy tries to match facial features such as eyes, nose, mouth, and chin to orient the head pose position.

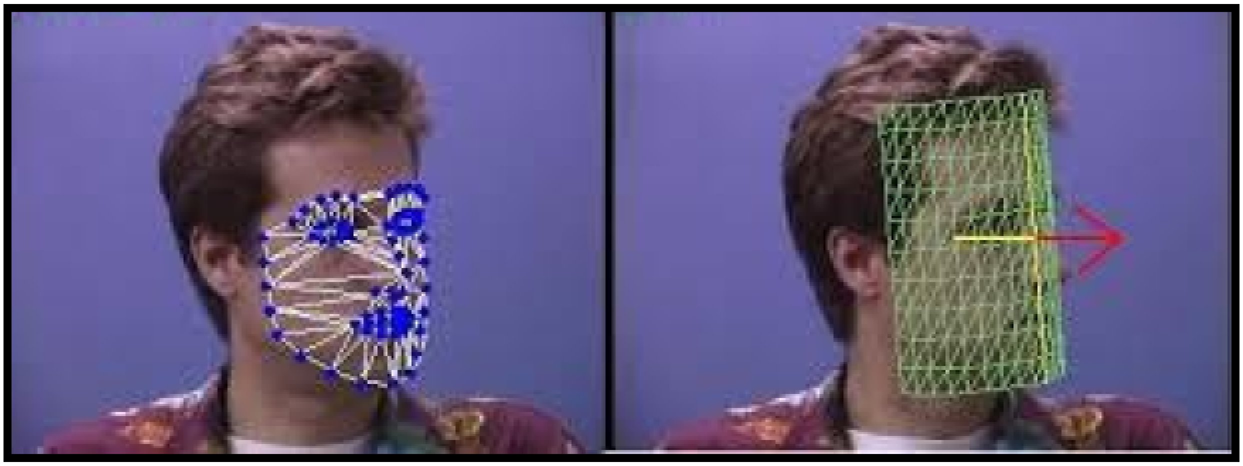


Figure 4: Face mesh matching for head pose [Sung et al., 2008].

In the concluding remarks, the survey proposes that full range pose estimation, even when the head is pointed away from the camera, should be one of the design criteria for future developments in this field of computer vision. This is a critical feature for Human-Robot team search missions since aft point of view of teammates is a common orientation scenario. One rare domain for aft head pose estimation, as seen in Figure 5, is security surveillance applications such as found in [Benfold, and Reid, 2009]. However, these security camera positions are typically from an elevated perspective, only provide coarse pose estimation, and focus on head pose direction in the ground-plane. These characteristics are not appropriate for HPT4M implementations.

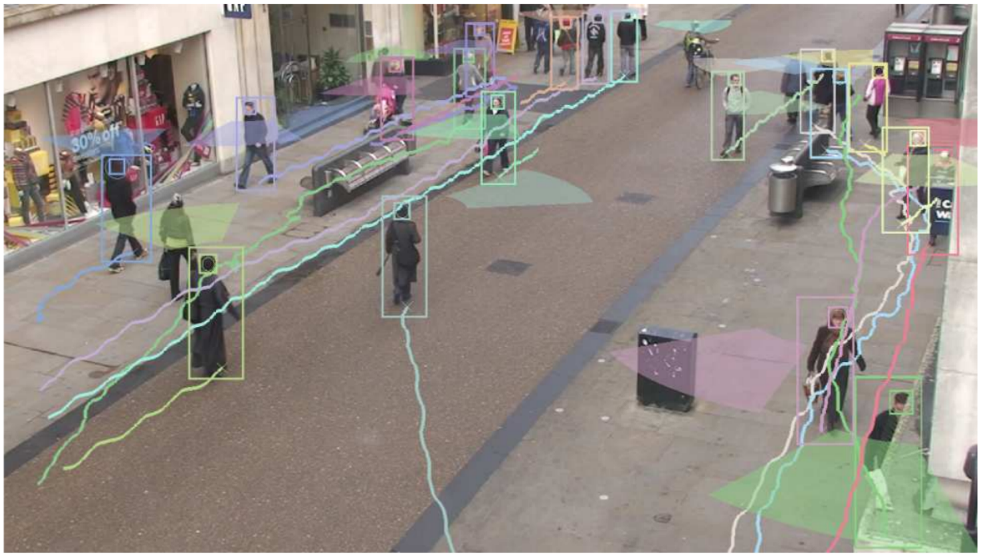


Figure 5: Aft head pose estimation from a surveillance camera [Benfold, and Reid, 2009].

Lastly, [Cao et al., 2017] present a modern approach for estimating real-time 2D human pose in image frames based on part affinity fields. The model fitting is based on eighteen body key points and is time invariant to the number of subjects detected. Although this approach allows for skeletal orientations to be extracted from 2D image streams in real time, converting to 3D skeletal pose requires simultaneously processing two stereo image streams. In practice, processing frame rates fast enough to match the thirty frames per second for two camera channels was found to be impractical. For example, a NVIDIA GeForce GTX-1080 GPU was only able to reach 8.8 frames per second in [Cao et al., 2017]. Furthermore, this approach still suffers from lack of subject detail from the aft perspective, so the potential pose accuracy in this scenario is unclear. Figure 6 shows an example of skeleton matching in OpenPose. None of the surveyed image processing techniques for this work provide fine head pose tracking from the aft point of view. Thus, another solution is needed for this thesis.

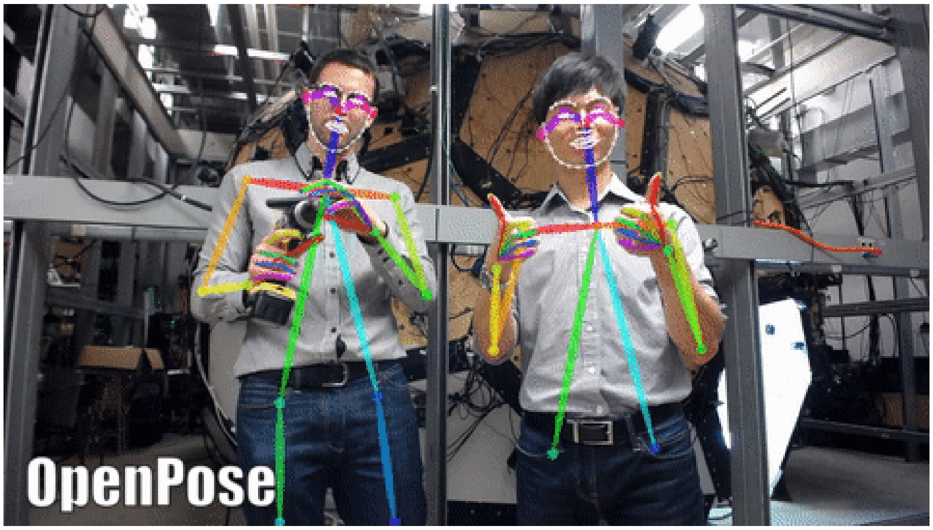


Figure 6: Pose estimation using part affinity fields [Cao et al., 2017][2].

2.3.2 Multi-Camera Motion Capture

Multi-camera motion capture systems are common for 3D human skeleton model extraction for use in movie special effects, video games, and studies of biomechanics. The systems often require special motion capture suits and an array of multiple cameras. An example of the multi-camera motion capture can be seen in Figure 7. Subfigure 7-A shows golf swing motion capture for video games, 7-B and 7-D show special effects motion capture for the 2009 film Avatar, and 7-C shows an example of a dog motion capture suit for use in biomechanics analysis.

The example images from Figure 7 require an array of multiple cameras in order to triangulate the tracking suit marker locations. Multi-camera tracking systems are not typically appropriate for real-world robotic applications due to this characteristic. This apparatus was not a valid candidate for this implementation of HPT4M because it does not meet the requirement to operate from a single perspective.

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/media/pose_face_hands.gif

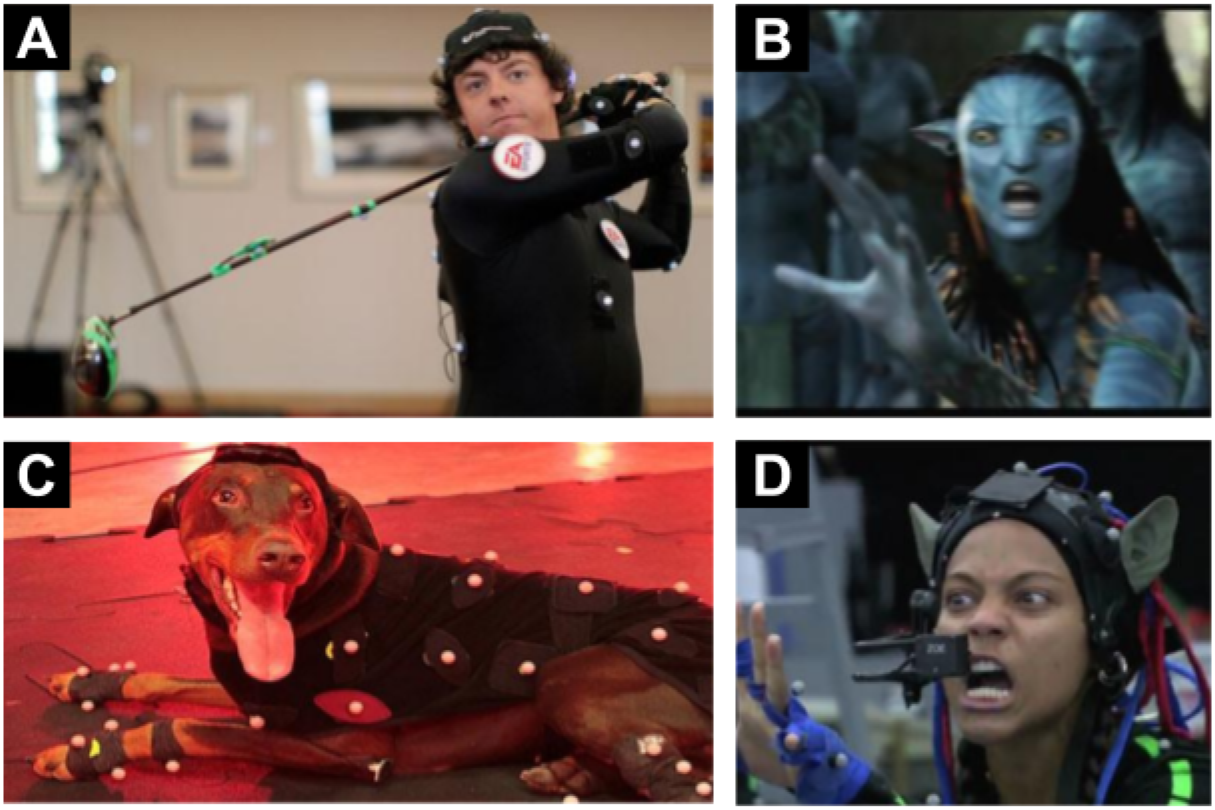


Figure 7: Multi-Camera Motion Capture Examples[3].

[3] Images sampled on 6/26/2017.
A: Golfer Rory McIlroy taking part in a motion capture session.
www.theguardian.com/technology/2010/jun/13/sports-sims-motion-capture-technology .
B&D: Avatar movie: www.james-camerons-avatar.wikia.com/wiki/File:Zoe_behind_the_scenes_2.jpg .
C: EA Capture Lab: www.engadget.com/2014/07/14/motion-capture-explainer .

2.3.3 Natural Interface Sensors

Another common system for 3D human pose capture, as discussed in [Zanuttigh et al., 2016], is natural interface sensors. As shown in Figure 8, these devices are commonly used for video game interfacing [Microsoft, 2017] and are considered a natural interface since they do not require the use of a special controller. More information on human pose extraction from RGB-D sensors can be found on the OpenNI open source project [OpenNI, 2017]. Although this approach works from a single sensor perspective, the skeleton extraction techniques used focus on the human body and limb positions and do not focus on the head pose orientation. The system assumes the user is observing the television monitor and thus high fidelity head pose is not required. Figure 9 shows that the extracted skeleton is basically headless, which leaves the desired head pose signal information absent. These characteristics fail to meet the given requirements for the HPT4M tracking systems.

Figure 8: Microsoft Kinect video game interface[4].

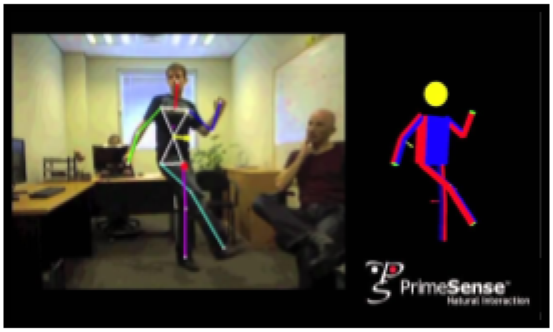


Figure 9: Skeleton extraction from natural interface sensor[5].

[4] Microsoft Kinect: www.developer.microsoft.com/en-us/windows/kinect .

[5] PrimeSense: https://i.ytimg.com/vi/nr8vgCnb9_0/maxresdefault.jpg .

2.4 Summary

Chapter 2 has discussed related work in the applicable areas of implicit communication for Human-Robot search team missions. This discussion supports the motivation in this thesis for deeper investigation of implicit human attention modeling in 3D space. Consequently, Section 2.2 discusses common techniques for modeling human attention and the inspiration for the HPT4M approach used in this thesis. HPT4M is a 3D extrapolation of the 2D concepts used in TMEG. The success of the TMEG human attention models in the domains of psychology, computer interfaces, and marketing leads to the conclusion that HPT4M is a valid technique to explore for use in Human-Robot team applications. It also drives the need for a head pose tracking system to compute the 3D attention map.

The survey of common human pose tracking systems concludes with the assertion that a custom head pose tracking solution must be implemented in order to investigate the hypotheses in this work. Fine head pose tracking from the aft perspective is not yet commonly available. Several common human pose tracking systems were investigated, but found to be inapplicable for HPT4M technology in this thesis, since they did not meet the system requirements defined in Section 1.1.

3 Hardware/Software Approach & Validation

This chapter discusses the system design, hardware, software algorithms, and validation approach for this implementation of HPT4M. This technology and validation testing were necessary to investigate the primary hypotheses of this research (as stated in Section 1.2):

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs.

The materials consist mainly of the RGB-D camera, the computer running the software, and a helmet marker apparatus for the human subject's head. The methods include two main HPT4M algorithms: the vision processing pipeline and the thermal mapping technique.

For validation, the tracking system's accuracy is first evaluated by measuring the reported head pose angles against ground truth sensors. The thermal mapping is then evaluated by rendering against a subject's attentional targets. Lastly, human tracking is evaluated by a bench top OOI search.

3.1 System Design

Figure 10 shows the HPT4M system design, which consists of a RGB-D camera, a robot system with a computer running the software, and a scene with a subject wearing the helmet apparatus. The simplicity of the HPT4M system design allows it to easily be used on different style robotic platforms such as humanoids and ground vehicles. As shown in Figure 11, this implementation targeted a Meka M3 humanoid hybrid ground vehicle system known as 'Rosie'. The system has two compliant arms, a mobile wheelbase, an audio system, and a RGB-D sensor mounted inside a head unit.

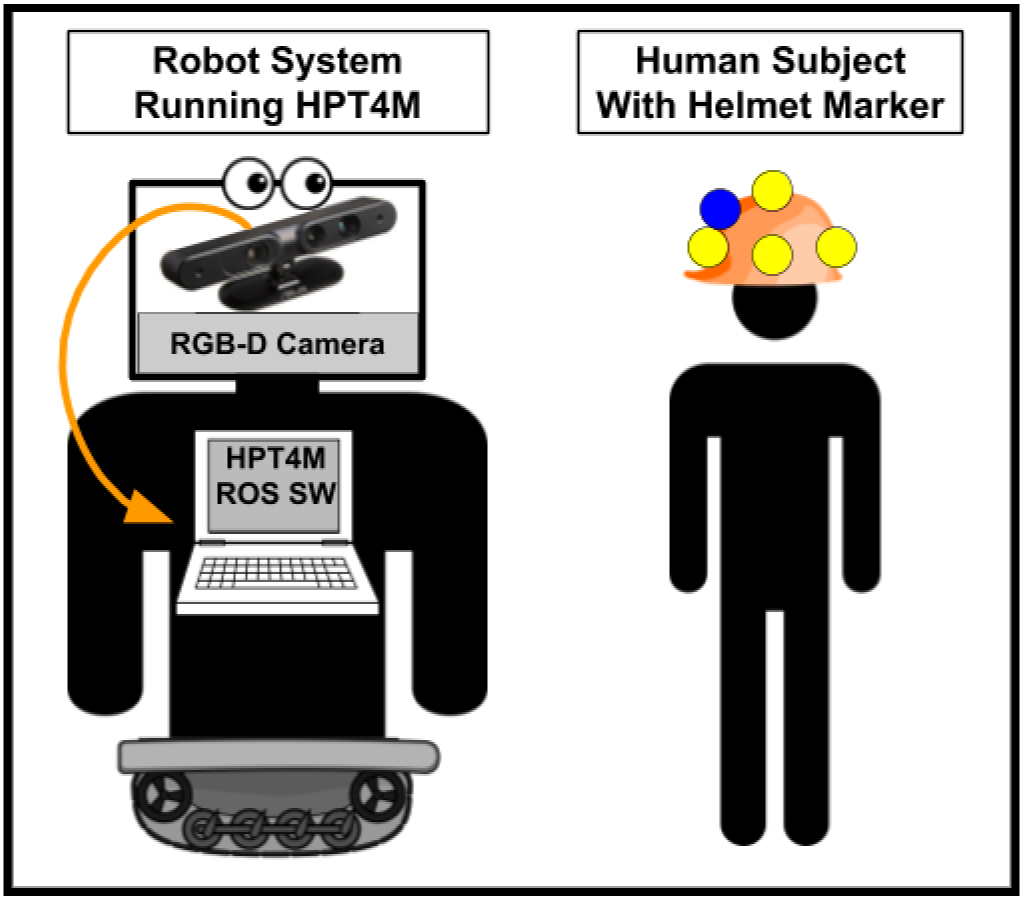


Figure 10: System Design.

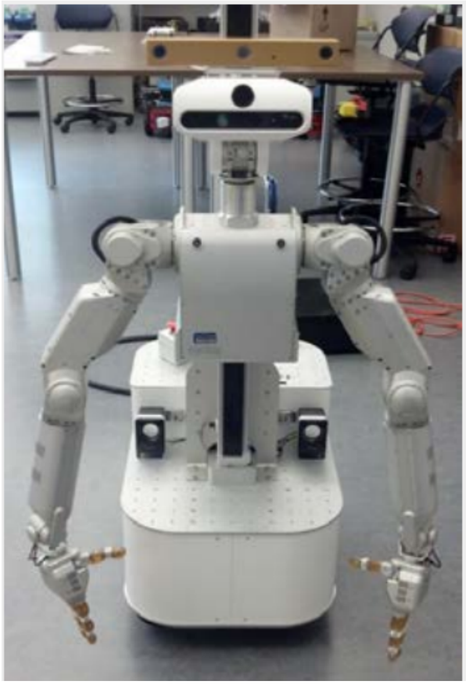


Figure 11: Meka model M3 robot.

3.2 RGB-D Camera

Chiefly, the HPT4M technology requires a point cloud scan from the agent perspective that contains a view of the subject's head; this was obtained by use of an RGB-D camera. This type of camera uses an active infrared light sensor to measure the depth of the environment and provide this data in an image frame. As Figure 12 shows, the ASUS Xtion PRO used in this work simultaneously captures the color and depth images. The RGB frame and depth frame information can be combined to form a 3D cloud of points by applying the depth frame pixels as a third dimension to the 2D RGB pixel coordinates (see Figure 13). More information on RGB-D cameras and applications can be found in [Zanuttigh et al., 2016]. The ASUS Xtion PRO model is no longer manufactured, but the modern Kinect depth sensor currently lists for $99.99 USD [Microsoft, 2017].

3.3 Helmet Marker Apparatus

This thesis requires the implementation of a custom head pose tracking solution. Section 1.1 lists the system requirements that drive the need for 360° head tracking and operation from a single sensor perspective. Also, the summary of Chapter 2 states why none of the typical human pose tracking systems are sufficient for these requirements. While we anticipate that video-based head pose trajectory tracking without special markers will soon be available, the custom head pose tracking system is developed in the meantime to enable us to explore the temporal thermal maps, which is the objective of this research. Again, the goal for the custom head pose tracking system is to support the investigation of the key hypotheses of this research, namely that:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs.

The helmet marker apparatus supports this effort by providing the HPT4M system with a standard head pose tracking target.

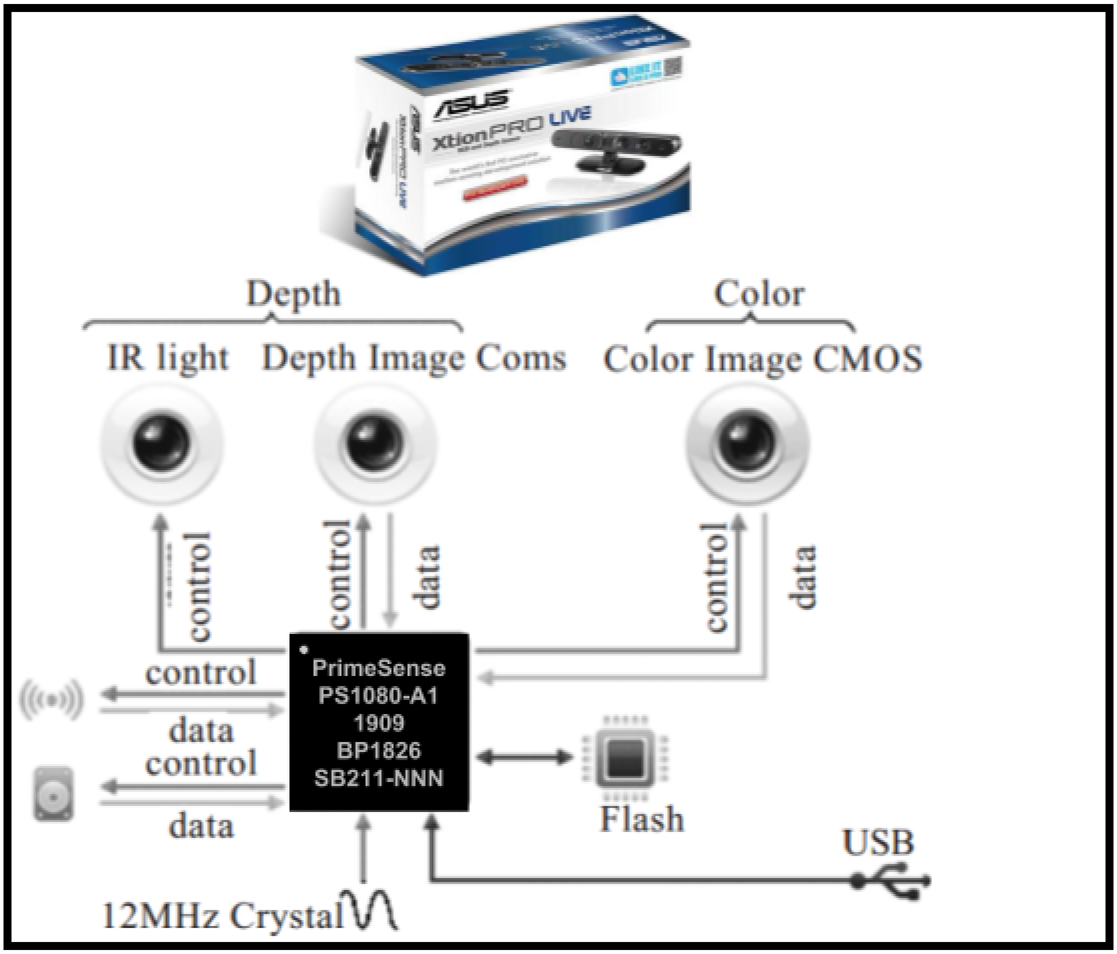


Figure 12: ASUS Xtion PRO Live RGB-D sensor and PrimeSense reference design[6].

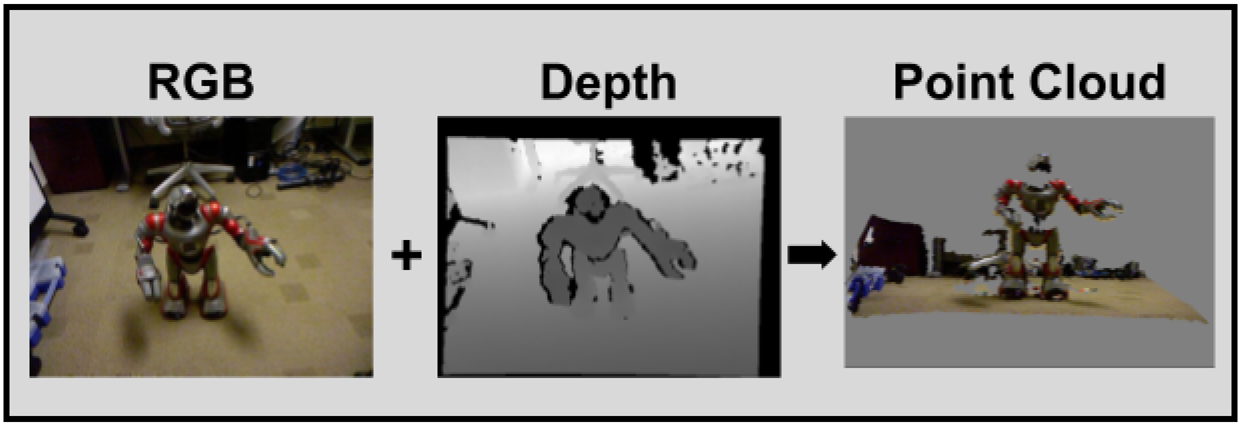


Figure 13: RGB & depth image combined into point cloud[7].

[6] ASUS Xtion PRO Live: www.asus.com/us/3D-Sensor/Xtion_PRO_LIVE/ .

[7] RGB-D camera images: www.wiki.ros.org/depth_image_proc .

Natural human heads are difficult to track accurately due to shape variance and lack of detail on the back side of the head. This work uses a marked helmet target, which facilitates fast, accurate, and consistent tracking of a subject's 360° head pose from a single RGB-D camera perspective. The apparatus consists of a standard construction-style safety helmet with tennis balls attached as seen in Figure 14-B. This design is chosen because the bright helmet and tennis balls are easy targets to localize and orient in 3D space based on a RGB-D camera data stream. Additionally, the helmet was chosen because it includes an occipital bone adjustment strap that allows the helmet to adapt securely to most human head sizes. Also, basic models can be easily found for less than $30 USD.

The helmet is marked with 6 tennis balls, which can be seen in Figure 14-B. Four of the yellow tennis balls surround the brim of the hat at 90° increments, starting with one at the forehead location. One blue ball marks the face side of the helmet and one of the yellow balls is located on the top crown. This pattern was chosen because it allows for three balls to be visible from most pose orientations and for the front side to be differentiated from the rear. Figure 14-A shows the head pose coordinate system being correctly targeted in a 3D point cloud.

A potential alternative design is a helmet marked with fiducial tags shown in Figure 14-C. Image processing libraries such as ArUco [Garrido-Jurado et al., 2014] can localize these tags in 3D space. However, this method has tracking difficulty due to camera resolution, lighting limitations, frame rate, and motion blur. As discovered in this work, the ball markers present a much simpler tracking target and are more robust against motion blur than the digital tags.

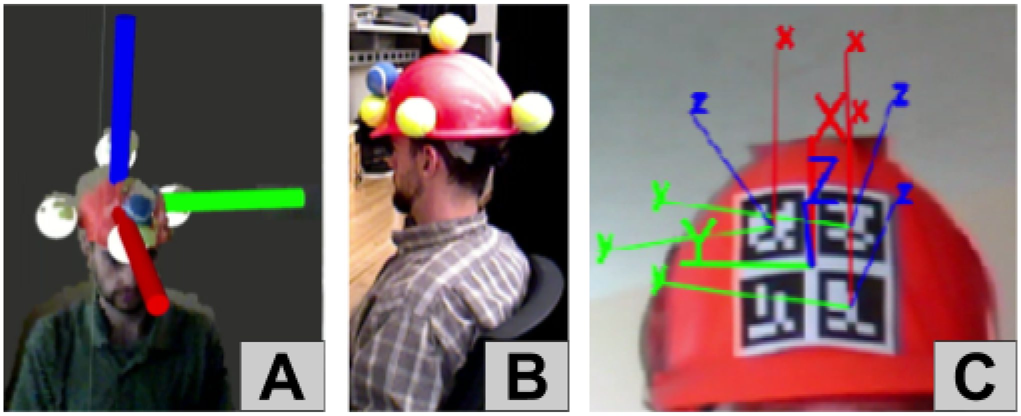


Figure 14: Helmet marker apparatus.

3.4 Algorithms

The HPT4M technology uses two key algorithms for thermal mapping and head pose tracking. The thermal map requires the head pose within the point cloud coordinate system. This information is used to project the head pose into the cloud point scene and render the thermal map as seen in Figure 15. The head pose tracking algorithm is covered in Section 3.4.1 and the thermal mapping algorithm is covered in Section 3.4.2.

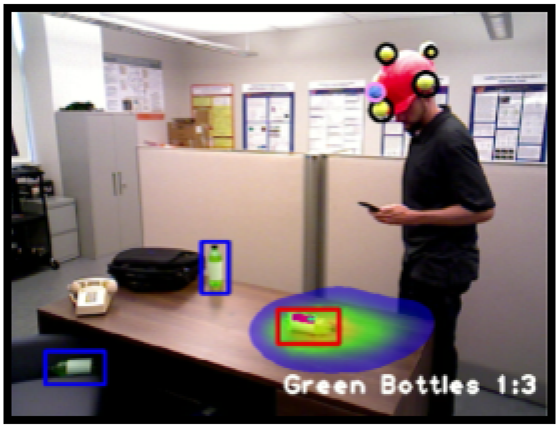


Figure 15: Thermal map rendered from head pose.

3.4.1 Head Pose Tracking

The algorithm for head pose tracking involves a multistage image processing pipeline as shown in Figure 16 and described in Figure 17. The bright orange helmet is first located in the full image frame (Fig. 16-A), which allows for extracting a hat Region of Interest (ROI) sub-image (Fig. 16-B) that contains fewer pixels. This smaller image is then checked for the yellow tennis balls. This allows for a periscope filter (Fig. 16-C) to be applied to the hat ROI, which helps block out further noise from the background such as the blue posters. Once this is complete, the blue hat ball can be located such that all available ball center coordinates are captured from the 2D RGB image frame (Fig. 16-D). Due to the depth sensor's functionality, any 2D point in the RGB image can be mapped to the corresponding 3D point cloud coordinate. This allows for collecting the 3D constellation of hat balls' coordinates, which can be correlated directly to head pose (Fig. 16-E) based on the coordinate geometry. Corrupt values are filtered out and any impossible pose frames are discarded.

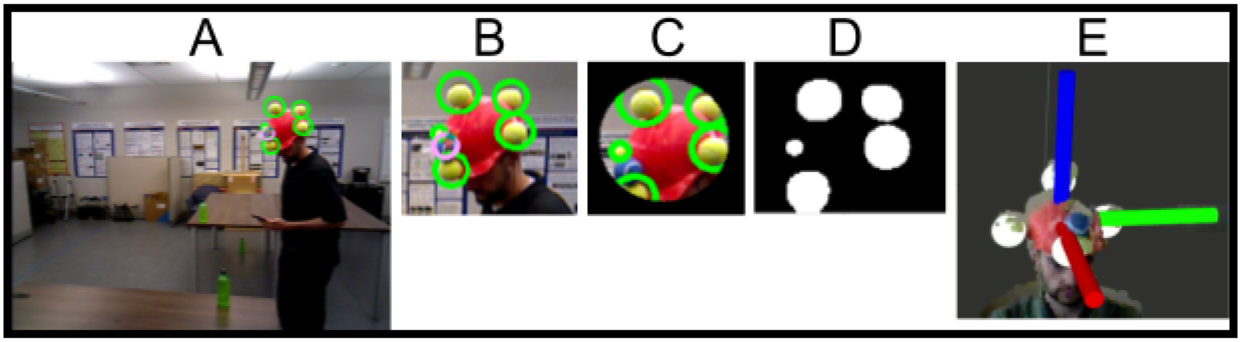
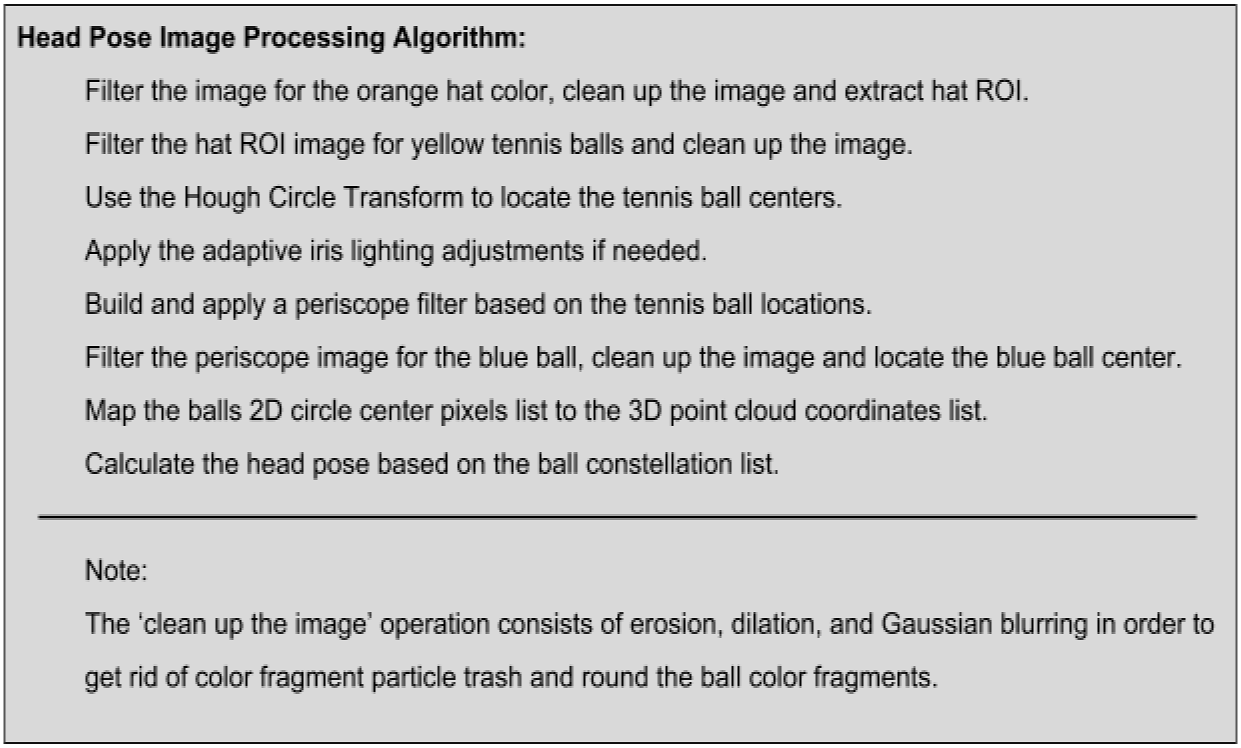


Figure 16: High level overview of helmet pose tracking pipeline.



Head Pose Image Processing Algorithm:

Filter the image for the orange hat color, clean up the image and extract hat ROI.

Filter the hat ROI image for yellow tennis balls and clean up the image.

Use the Hough Circle Transform to locate the tennis ball centers.

Apply the adaptive iris lighting adjustments if needed.

Build and apply a periscope filter based on the tennis ball locations.

Filter the periscope image for the blue ball, clean up the image and locate the blue ball center.

Map the balls 2D circle center pixels list to the 3D point cloud coordinates list.

Calculate the head pose based on the ball constellation list.

Note:

The 'clean up the image' operation consists of erosion, dilation, and Gaussian blurring in order to get rid of color fragment particle trash and round the ball color fragments.

Figure 17: Head pose image processing pipeline algorithm.

3.4.2 Thermal Map Projection

The thermal map projection makes use of the acquired head pose and point cloud information. The head pose is placed in the point cloud coordinate system and the gaze projection is cast into the point cloud. As shown in Figure 18 and described in Figure 19, points contained within the gaze are colorized by a thermal induction algorithm that heats up the points if they are within the gaze envelope and allows them to cool off otherwise. Points more central to the gaze pattern are induced with a greater level of heat and thus heat up more quickly than points located closer to the gaze periphery.

3.4.3 Parameter Settings

The thermal mapping algorithm has several parameters to set the runtime execution behavior. These parameters include the thermal map induction heat rate, cool off rate, and the gaze envelope diameter. In this work, symmetrical thermal rates are chosen to model full human attention within three seconds of gaze time. Alternatively, dynamic thermal rates could be implemented to adapt to different recognized behaviors representing varying levels of attention. For example, recognizing sitting vs. searching behaviour could drive dynamic thermal rates to more appropriately reflect these modes of attention. However, this complex concept is considered out of scope for this initial exploration of HPT4M.

The gaze envelope diameter for this work is selected as one-half meter because of its effectiveness at close range experiments of approximately three meters from subject to target OOI. An exploration of an anatomically based gaze envelope design is a strongly recommended candidate for future explorations of HPT4M implementations. Additionally, eye gaze is assumed to match head pose in this work. Further investigation into the complex impacts of eye gaze in relation to head pose are also considered out of scope for this initial exploration of HPT4M. Obtaining this information would require real time eye tracking of a subject’s gaze while in motion in 3D space.

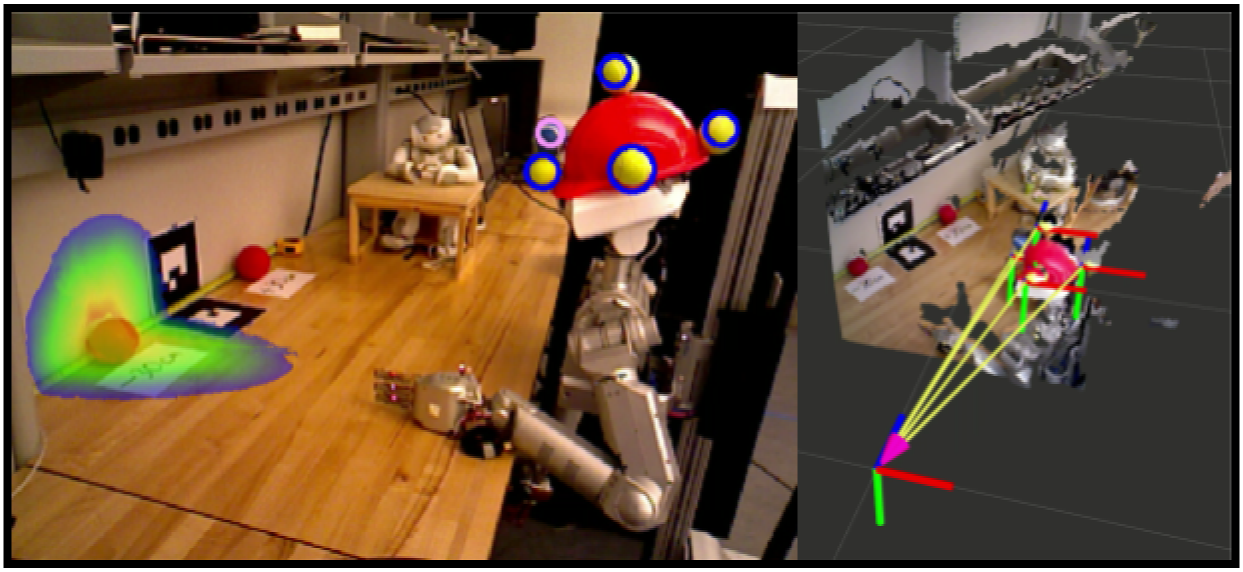
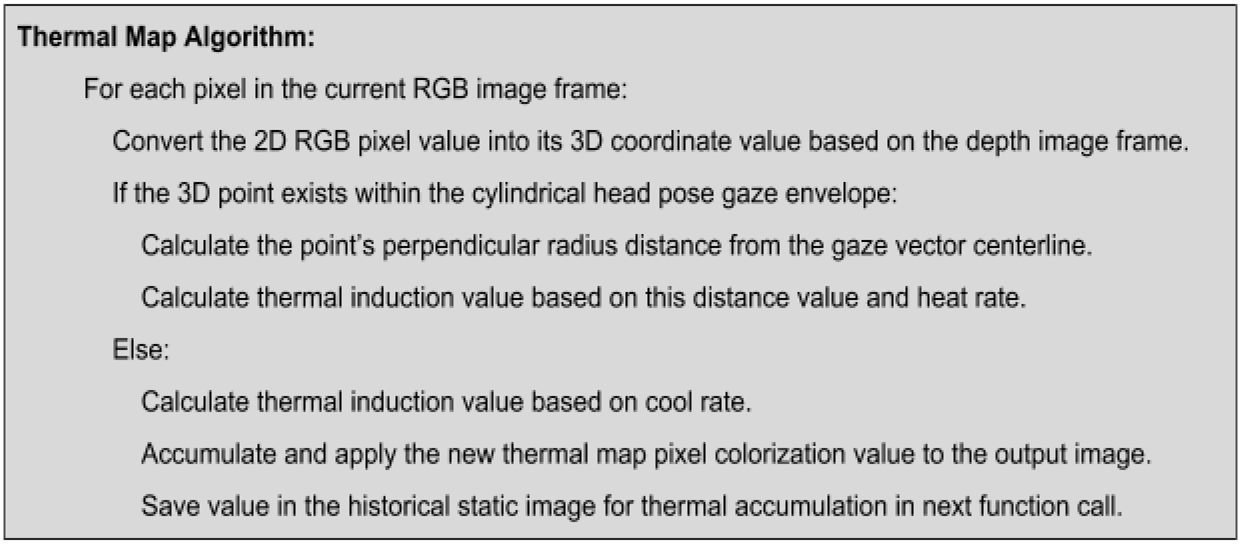


Figure 18: Temporal thermal map generated from head pose.



```
Thermal Map Algorithm:
    For each pixel in the current RGB image frame:
      Convert the 2D RGB pixel value into its 3D coordinate value based on the depth image frame.
      If the 3D point exists within the cylindrical head pose gaze envelope:
        Calculate the point's perpendicular radius distance from the gaze vector centerline.
        Calculate thermal induction value based on this distance value and heat rate.
      Else:
        Calculate thermal induction value based on cool rate.
        Accumulate and apply the new thermal map pixel colorization value to the output image.
        Save value in the historical static image for thermal accumulation in next function call.
```

Figure 19: Thermal map algorithm.

3.4.4 Object of Interest Detection

This work use two types of image classifiers for OOI detection. The first is a simple classifier that can distinguish between red balls and green bottles based on the object's shape and color. This method uses OpenCV [Bradski, and Kaehler, 2000] which has the advantage of executing quickly and efficiently. The second technique for object classification allows for a larger and more general set of objects to be classified by using TensorFlow [Abad et al., 2015] for inference on a specially trained model. The Inception-V3 model [Szegedy et al., 2015], which was originally trained on Imagenet [Russakovsky et al., 2009], was modified by using TensorFlow to retrain the final layer with a set of training images taken from common laboratory objects used in in this work.

The left image of Figure 20 shows an example of the HPT4M system highlighting the red balls as the subject's Objects of Interest. The right image of Figure 20 shows a red ball extracted in a ROI subframe. The red balls are detected based on color range and roundness, which is detected by the OpenCV Hough Circle Transform. Green bottles are also detected by color range but the OpenCV contours method is used to derive a bounding box on the green fragments, with the height to width ratio of the long bottles used for detection.

An alternative approach for object classification would be to use a remote classifier such as such as the 'Google Cloud Vision API' [Google, 2017]. This would allow for a continually up to date and diverse object classification capability, but it has the penalty of delayed response and can require data limits or processing fees. The Google image classifier provided in the 'Cloud Vision API' has impressive results, as seen in Figure 21. However, the confidence levels for these simple objects still remain dispersed among several unrelated items such as 'jewelry' for the red ball. The simple OpenCV based classifier developed for this work was found to be sufficient for the initial experiments. The custom TensorFlow inference model was used for the more challenging clustered OOI discrimination experiments.

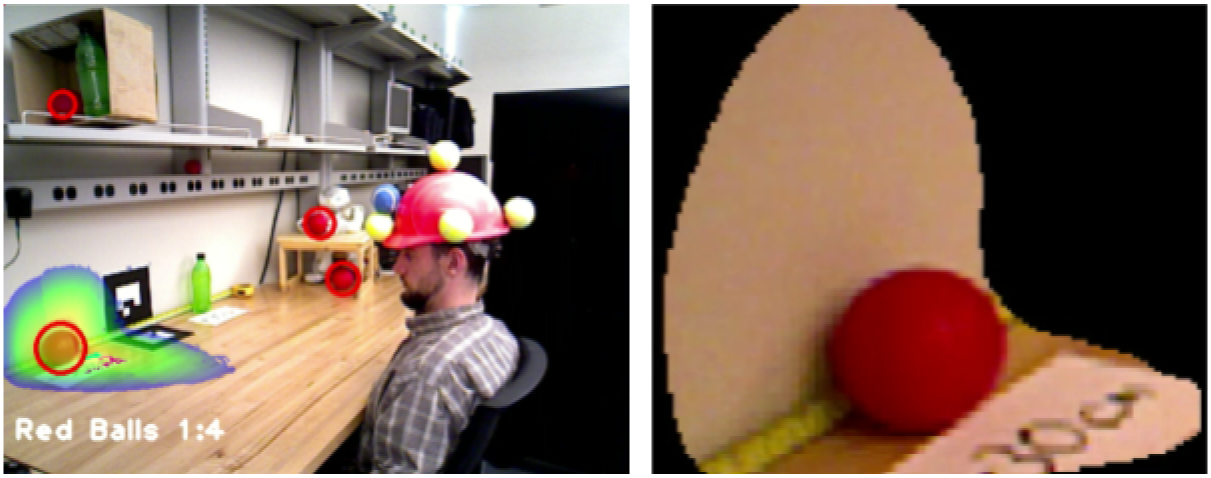
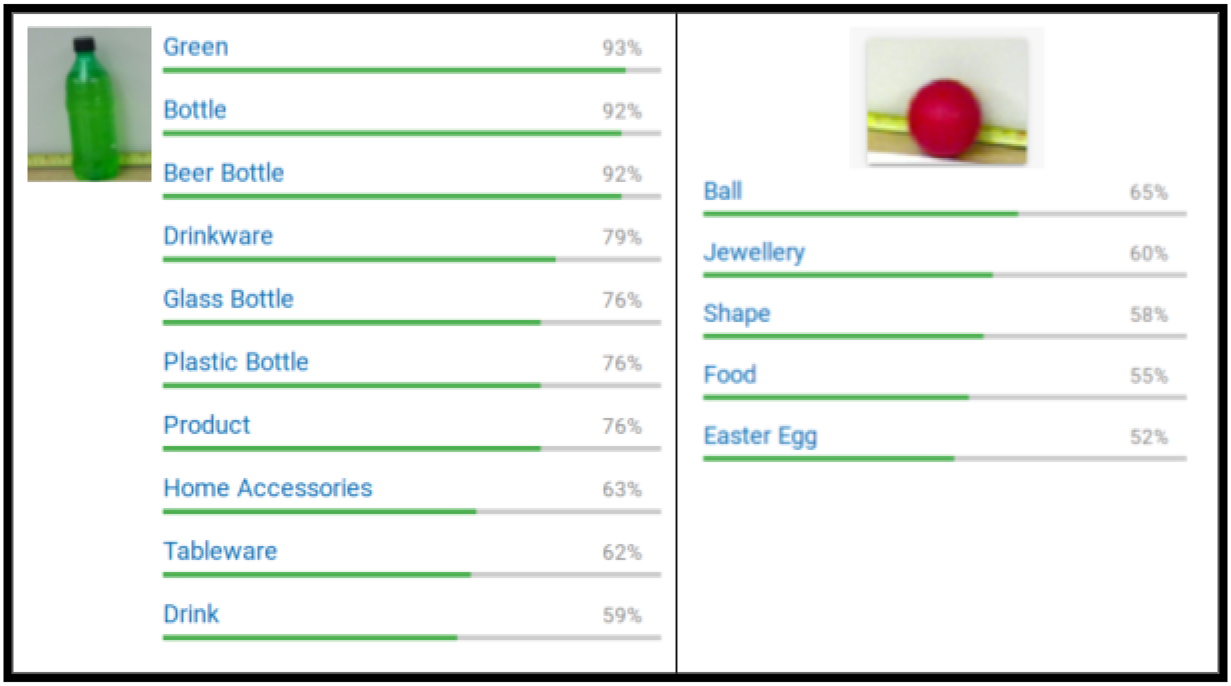


Figure 20: Human subject Object of Interest detection and extraction.



Green	93%
Bottle	92%
Beer Bottle	92%
Drinkware	79%
Glass Bottle	76%
Plastic Bottle	76%
Product	76%
Home Accessories	63%
Tableware	62%
Drink	59%

Ball	65%
Jewellery	60%
Shape	58%
Food	55%
Easter Egg	52%

Figure 21: Google image classification example[8].

[8] Google Cloud Vision API: cloud.google.com/vision/ .

3.5 Robot Operating System

The Robot Operating System (ROS) [Quigley et al., 2009] is selected as the software node framework for the HPT4M implementation. ROS allows for standardized message passing among software node modules and is designed for collaboration among different robotics research groups. It also has numerous development tools such as visualizers, message recording, message playback, automatic node diagrams, and more. Designing this implementation of HPT4M with ROS allows for easy modular adaptation of the capabilities for use and testing in other systems in the robotics developer community and helps propagate the potential contributions of this work.

3.6 Verification of Head Pose Tracking

The first important evaluation for HPT4M is to measure basic head pose tracking accuracy against ground truth. All other experiments and evaluations depend on this core ability to acquire accurate head pose information from the subject. A Meka humanoid robot known as 'Rosie' was used as a stand-in measurable human proxy (see Figure 22). The Rosie head sensor unit contains an RGB-D camera plus motors and joint axis angle sensors for pitch and yaw. This apparatus provides ground truth information for tracking comparisons. Figure 22 shows the Point Of View (POV) from the Observing Agent (OA) camera on the left half and Rosie's camera POV on the right half. The red balls with the green circles in Figure 22 are the same ball as seen and recorded simultaneously from the two different camera points of view. The two fiducial markers act as a common home origin point for the two camera coordinate systems.

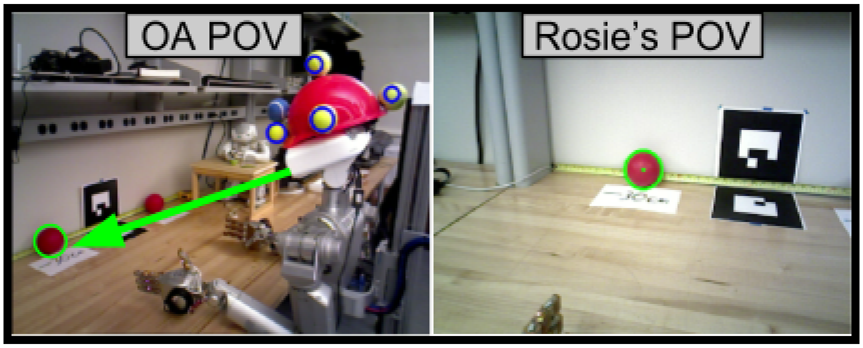


Figure 22: Apparatus to verify head pose tracking.

3.6.1 Successful Angular Sweep Test

The first type of validation test consisted of a simple periodic 30° angular sweep from center to head yaw left. The Observing Agent image feed was processed with the HPT4M pipeline and the reported head pose orientation was logged simultaneously with the neck axis angle. Because the two pose signals were captured over the same local computer network at the same time, the plots can be graphed together and overlaid for comparison.

The Figure 23 represents the 30° angular sweeps of the head yaw. The HPT4M system is able to track the head pose in this experiment with an average error of 1.62° and a standard deviation of 1.24°. This level of accuracy is attributed to the sensor's ability to accurately correlate the RGB pixels to the depth scan and is also demonstrated by the sensor's ability to render realistic point cloud scenes. These values were found to be within acceptable limits by confirming in the next experiment that the thermal map was able to land correctly on the subject's attention targets. Possible improvements for the error rates could include characterizing the signal noise and applying an output filter or upgrading to a higher quality depth sensor such as a modern Kinect camera.

3.6.2 Failed Angular Sweep Test

Figure 24 shows an early attempt of the angular sweep test. Originally, the software components were split into separate ROS nodes (see Figure 25) as is the standard approach for modular ROS development. However, this breaking up of modules caused messaging overhead, redundant image streams and delays in the transport layers. The poor performance results of this architecture are represented by the low granularity of the orange head pose tracking signal. Consequently, the software was reorganized to merge these nodes inside one master node, at the cost of losing software component modularity. This unification avoided the external transport delays between nodes and resulted in the successful tests in Figure 23.

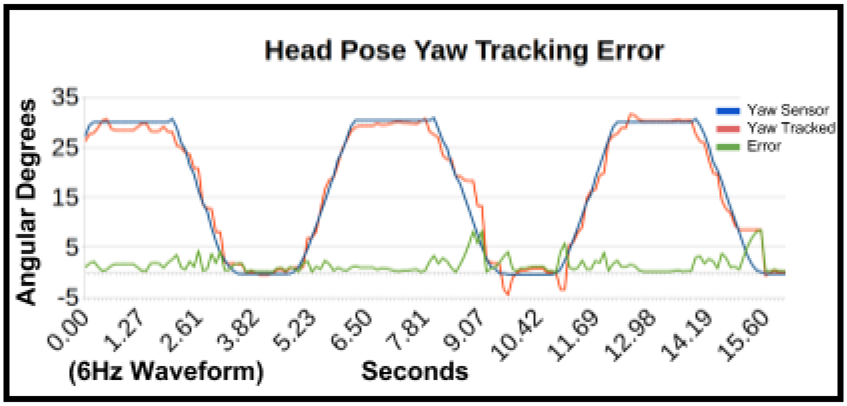


Figure 23: Graph of HPT4M head pose vs sampled head unit value.

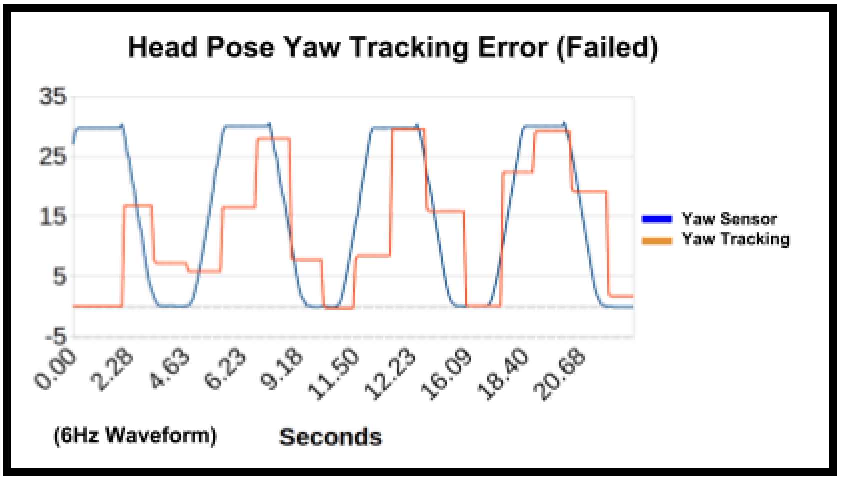


Figure 24: Graph of an early attempt at system verification.

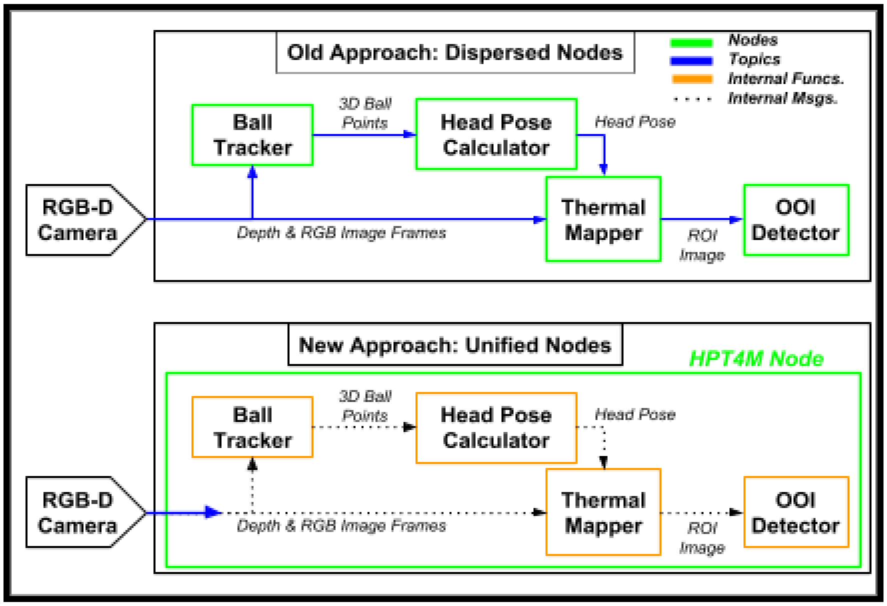


Figure 25: Software Nodes and Topics.

3.6.3 Thermal Map Validation Test

In the next type of verification test, Rosie was programmed to periodically move the head vision sensor unit to optically target the left or right red ball into the center of the image frame within a +/- 5 pixel window. Figure 26 shows the ball targeting frame from Rose's POV in the right half and the bench apparatus in the left half. Again, both camera feeds were captured over the same computer network, which allowed for simultaneous logging.

As seen in Figure 27, the optical controller tracking software was originally prototyped by using the Webots simulator [Cyberbotics, 2017]. This allowed for faster development time of the optical controller since the developing software on the hardware apparatus slows the process down. Simulations allow for quicker apparatus resets, easier system pause during development, and protect expensive hardware from early software errors.

Figure 28 shows four video frames from the optical controller thermal mapping verification test. The sweep transitions from targeting the left red ball in frame one to then targeting the right red ball. The thermal induction effect can be seen in the middle two frames where the thermal blob is fading from the left ball and ends on the right ball in frame four. The thermal map successfully lands centered on the two red balls, thus confirming the ability of the HPT4M system to accurately render onto the observed targets. The Rosie camera point of view allows for confirmation of the subject's targets, as can be seen in the top frames of Figure 28.

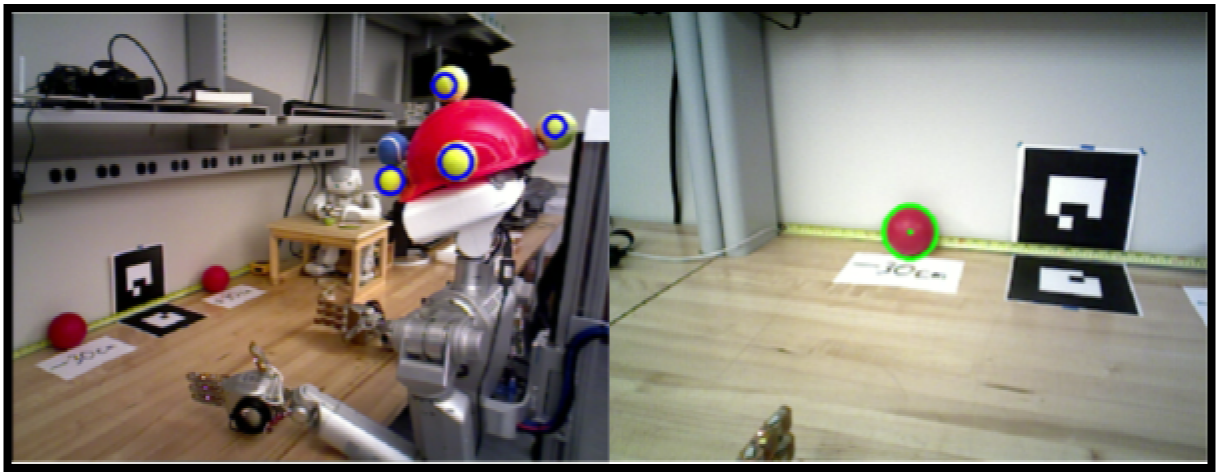


Figure 26: Apparatus to verify thermal mapping

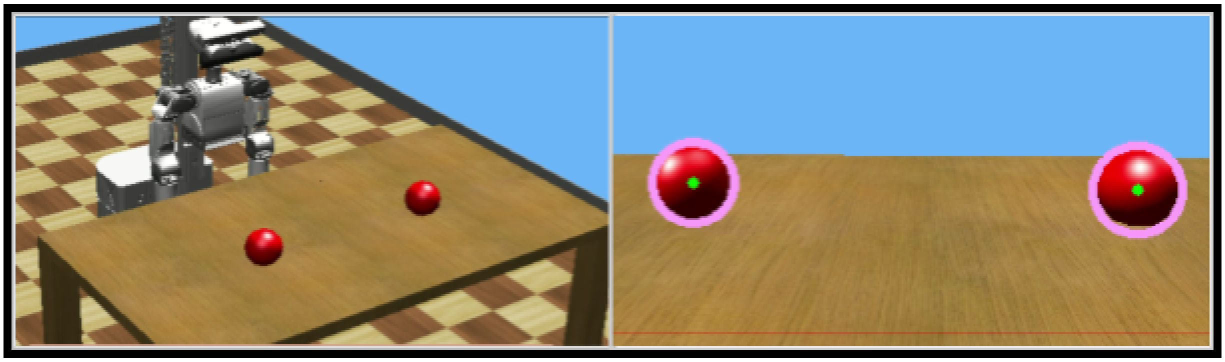


Figure 27: Optical controller simulation in Webots.

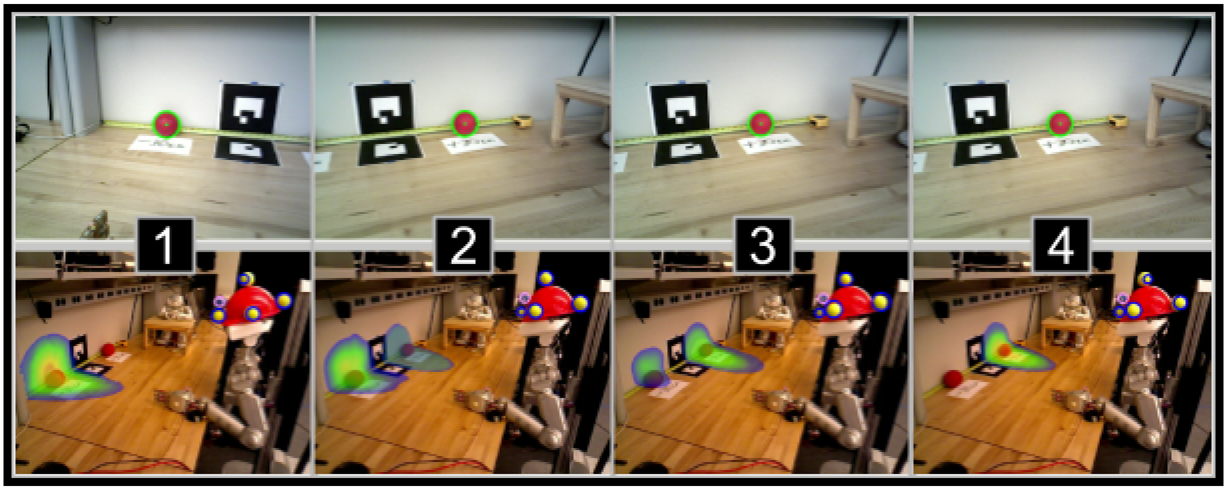


Figure 28: Four video frames from the HPT4M verification test.

3.6.4 Human Bench Tests

After the two types of verification tests with the robotic human proxy system, this bench apparatus was evaluated with a human subject before proceeding to the laboratory team search experiments discussed in Chapter 4. In this test, one of the red balls was replaced with a green bottle and extra objects were added throughout the scene. This allowed the Object of Interest detection software to be prototyped and evaluated as well. The green bounding boxes and red bounding circles represent the subject's current type of OOI as detected by the thermal map's extracted ROI object classification. The OA labels the OOIs and the ratio of the subject's discovered OOIs vs. total OOIs is displayed as text in the bottom left of the RGB frame. Of note in Figure 29 is that the items in the cardboard box are out of view and hidden from the subject. This showcases an example where the Observing Agent could potentially alert the teammate to the undiscovered objects.

In this experiment, the human subject targets the left and right objects with the same frequency as the humanoid robot. The HPT4M system is able to correctly detect when the subject focuses attention on the red ball and green bottle. These transitions cause other OOIs in the scene to be labeled and counted. This shows the potential for HPT4M to contribute to a team search mission without the need for explicit communication.

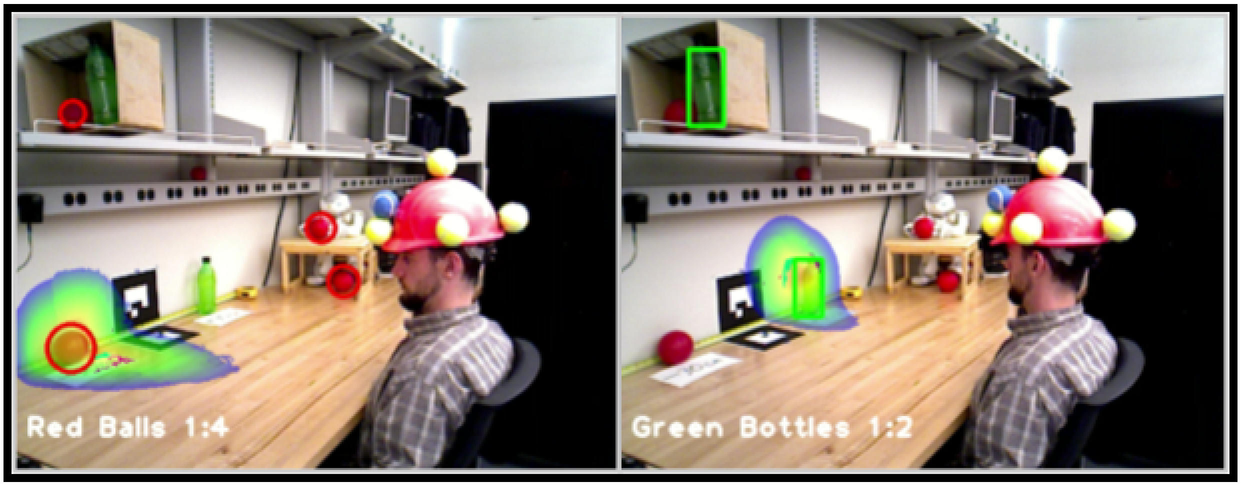


Figure 29: Human bench testing.

3.7 Summary

Chapter 3 shows by validation testing that the approach and design choices for the hardware and software in this thesis support the hypotheses by showing that the HPT4M attention model can accurately reflect a human subject's OOIs and contribute to team search by mapping these items locations. The included algorithms are simple in concept but require several image processing stages chained together in the correct architecture. Also, the custom Object of Interest detector results in faster operation than the cloud classification technique but also significantly limits discoverable types of OOIs. The selection of ROS for the software framework should allow for easy collaboration and sharing of the techniques in this thesis with the robotics community.

Additionally, the approach and design choices result in a low cost system with simple component architecture. The simplicity of this system and small sensor size make it flexible enough to work on many types of common robots. In addition, the head pose marker apparatus is cheap, durable, and accommodates a wide variance of human subject head sizes. This new head pose trajectory tracking tool allows the hypotheses in this thesis to be investigated with the experiments included in Chapter 4.

4 Experiments, Results, and Discussion

This chapter includes details, results, and a discussion on experiments that use temporal thermal mapping and alternative techniques in a Human-Robot team search mission. These experiments are in support of proving the posed hypotheses:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs

4.1 Team Search Experiment

This team search experiment is designed to investigate the posed hypotheses by determining whether HPT4M can be used in Human-Robot teams such that a human subject's Objects of Interest (OOIs) are implicitly discovered and mapped by a robot Observing Agent (OA). As seen in Figure 30, search mission examples where robots could be valuable team contributors include survivor recovery during disasters, hazardous materials spill cleanup, and search for objects of interest such as meteorites. Figure 31 shows the robot fleet of the Distributed Intelligence Laboratory of The University of Tennessee. Robots such as these are candidates for using HPT4M technology to assist with Human-Robot search missions.

Figure 30: Hazardous team search mission examples[9].

[9] Images retrieved on on 6/21/17:
Left: NASA Arctic meteorite search team.
earthandsolarsystem.files.wordpress.com/2013/01/recon-team-collecting-meteorite-at-szabo-bluff.jpg .
Center: FEMA team searching for tornado survivors.
www.fema.gov/blog/2013-05-23/oklahoma-tornadoes-update-photos-ground .
Right: Kansas Firefighters inspecting hazardous chemicals: www.firemarshal.ks.gov/division/hazmat .

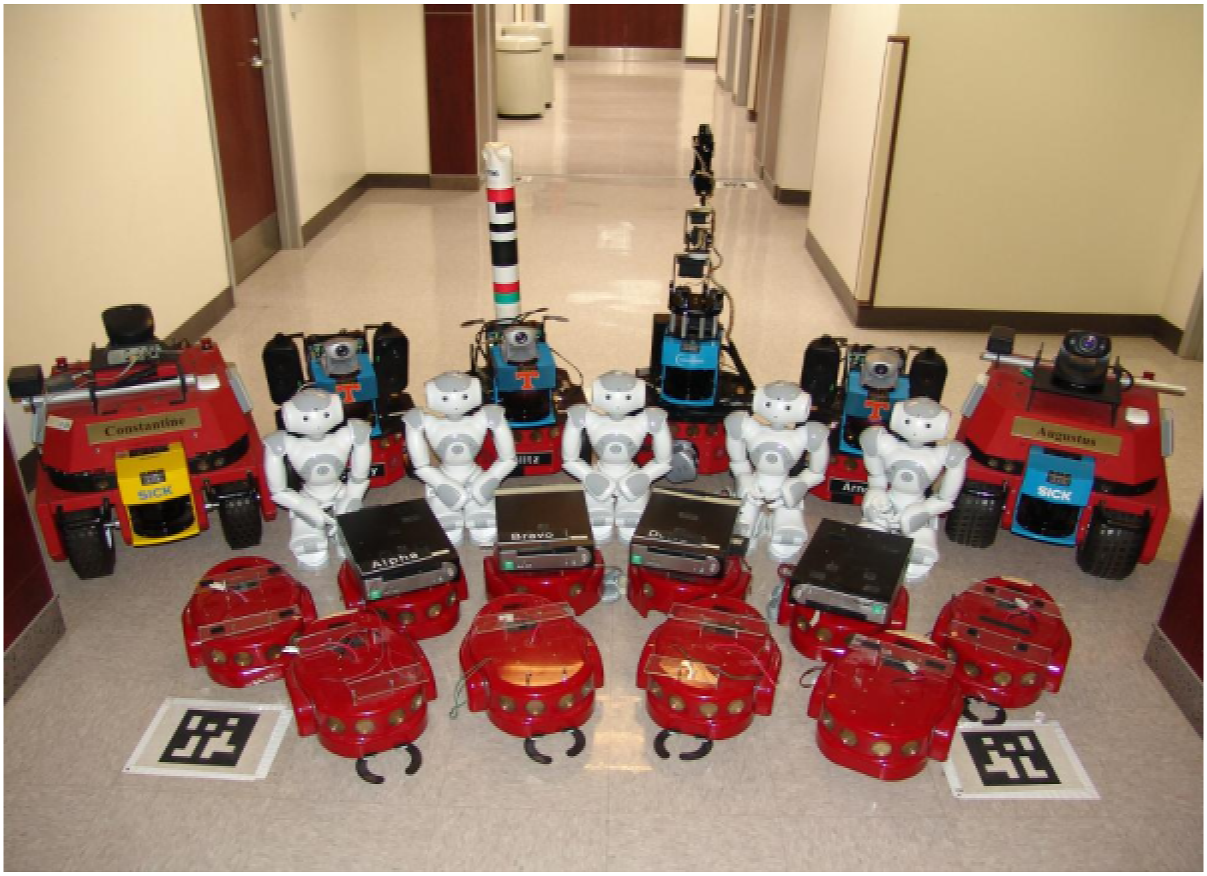


Figure 31: Potential future search team members[10].

[10] U. of Tennessee Distributed Intelligence Laboratory: http://dilab.eecs.utk.edu/DILab-robots.jpg

4.1.1 Team Search Experiment Overview

The team search mission is an experiment that involves one Observing Agent (OA) and one human search team participant. As shown in Figure 32, the Objects Of Interest (OOIs) in this experiment are a set of green soda bottles that are are placed in a simple experimentation room with the OA that runs the HPT4M software. The participant is asked to don the tracking helmet and then enter the room to explore for OOIs. The targets are found and scanned by the participant by taking an approximately three second video with a camera phone (ostensibly to measure and document the OOI). The experiment concludes once the subject determines that all the OOIs have been discovered and scanned, or once five total minutes have passed. During the experiment, the OA executes the HPT4M software that observes the human participant. Once the OOIs are determined, the OA highlights OOIs and indicates their location to the human subject.

Figure 33 shows a run of the experiment with the human subject. The OA builds the temporal thermal map based on the head pose trajectory as the subject moves in the scene. When the subject stops to meter the OOIs with the scanning device, this causes the OOIs to heat up in the map. The OA then realizes that the green bottle is the subject's OOI and begins to search for the items elsewhere in the scene. The OA also keeps a record of which items have been observed. For example, as marked by the OA in Figure 33, the green boxed object was previously observed, the red box indicates that bottle is currently being observed, and the blue box indicates an unobserved object that is blocked from the subject's view by the table and chair arm. If the subject exits the scene without scanning the blue boxed object, then the OA would have the information necessary to alert the human team member.

4.1.2 Team Search Experiment Results

The results of this experiment indicate that the HPT4M system is able to implicitly determine subject's OOI. The system is able to historically track which objects were observed and when the discovery event occurred. As shown in Figure 34, the first bottle is discovered and labeled red in video frame 1. In frame 2, the middle bottle becomes the current OOI so it is labeled with a red box and the prior OOI becomes green to indicate that it has already been observed. Note that in video frame 3, the hidden OOI remains labeled with a blue box because it was never discovered.

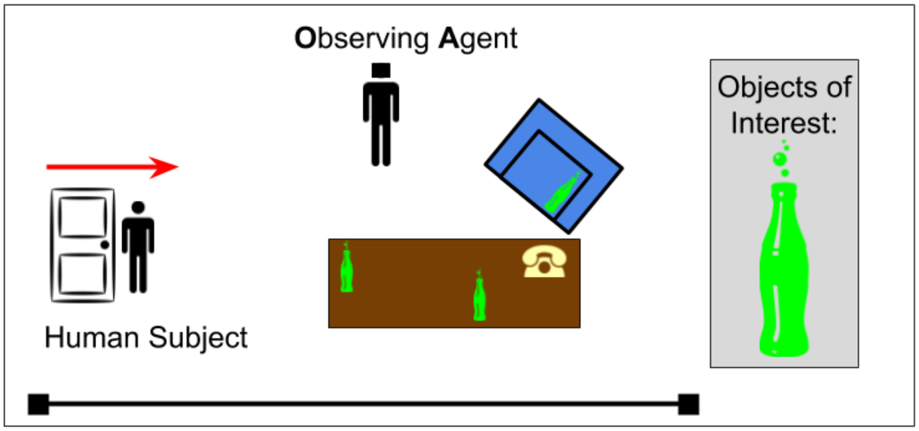


Figure 32: Search mission experiment diagram.

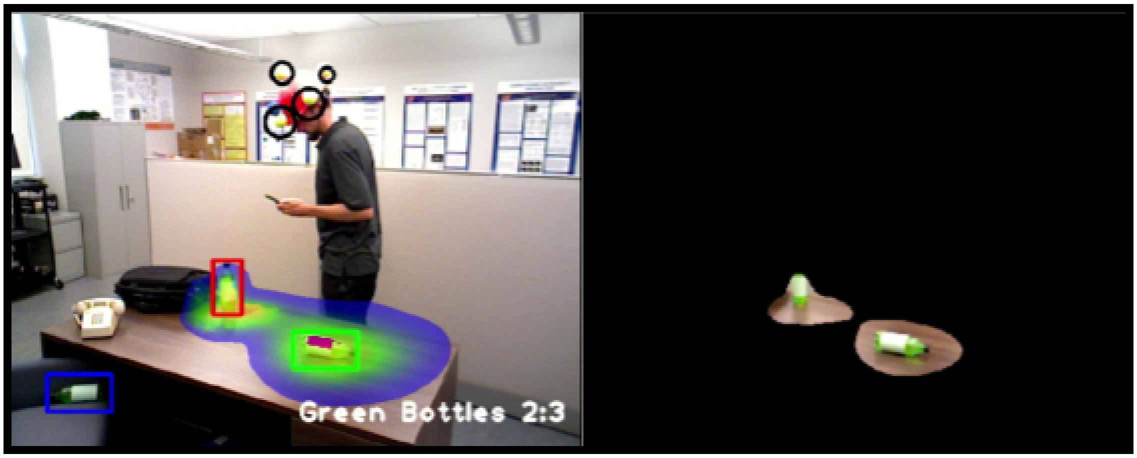


Figure 33: Search mission experiment.

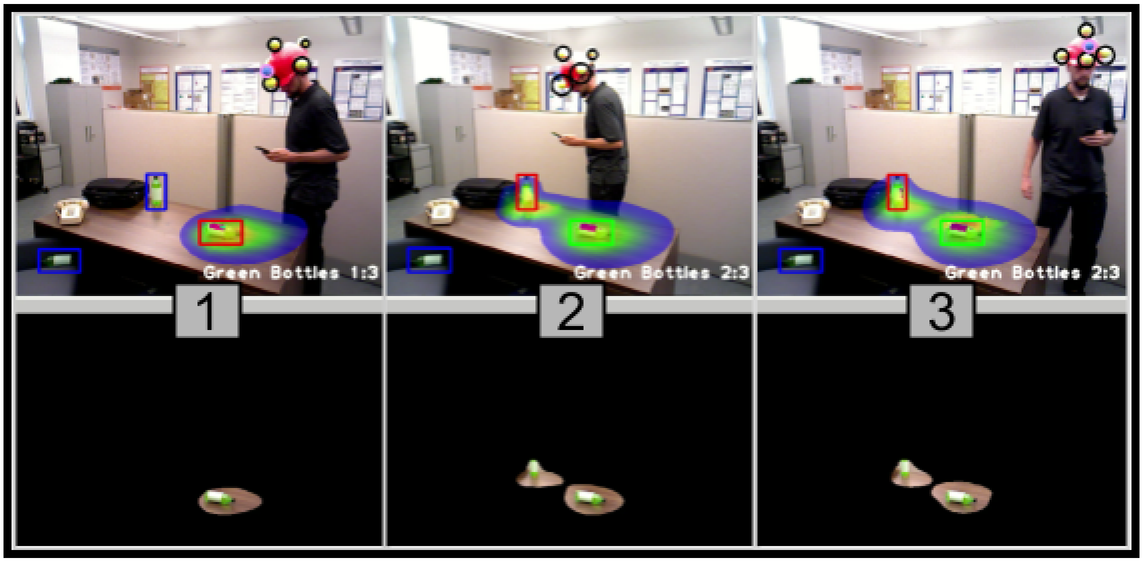


Figure 34: Search mission experiment sequence.

4.1.3 Team Search Experiment Discussion

This simple experiment shows an entry-level capability of HPT4M to contribute to a Human-Robot team search task by using implicit communication. The Observing Agent is able to passively determine the subject's OOI, highlight the OOIs, determine which ones are unobserved by the subject, and map the searched and unsearched space.

Future efforts in this work include OA audio and speech dialog capability for interacting with the subject. Currently, the subject has to observe a computer monitor to view the OA information; however, this interface is not practical for real-world scenarios and does not fit the long-term desired goal of efficient Human-Robot teamwork. In addition, the OA could use gestures to clarify OOI locations. For example, since the OOI 3D coordinates are known to the OA, this would would allow the OA to direct the subject to the unobserved objects by gesture pointing.

4.2 Alternatives to Temporal Thermal Mapping

A additional experiment was conducted to help evaluate HPT4M against alternative techniques that could be used in the team search task. First, a temporal non-thermal technique was implemented and is shown in the center column of Figure 35. This technique simply accumulates the gaze map over time and does not colorize it with the thermal induction calculations. This technique is able to build an accumulated spatial map of observed locations over time but it is not able to differentiate newer gaze attention locations from older ones within the map. Also, due to the absence of thermal colorization, it is not able to differentiate the central zones of attention from peripheral zones. Despite these drawbacks, the classifier is still able to successfully extract the green bottle OOI from the map's region of interest (ROI) in Figure 35-Y.

The left column of Figure 35 shows the non-temporal non-thermal implementation. This approach simply extracts the ROI instantaneously from head pose. It works in this experiment because the OOI classifier is able to extract the green bottle from the ROI. However, for situations where the subset of expected OOIs is unknown, this technique would not have the advantage of temporarily accumulating a spatial attention map.

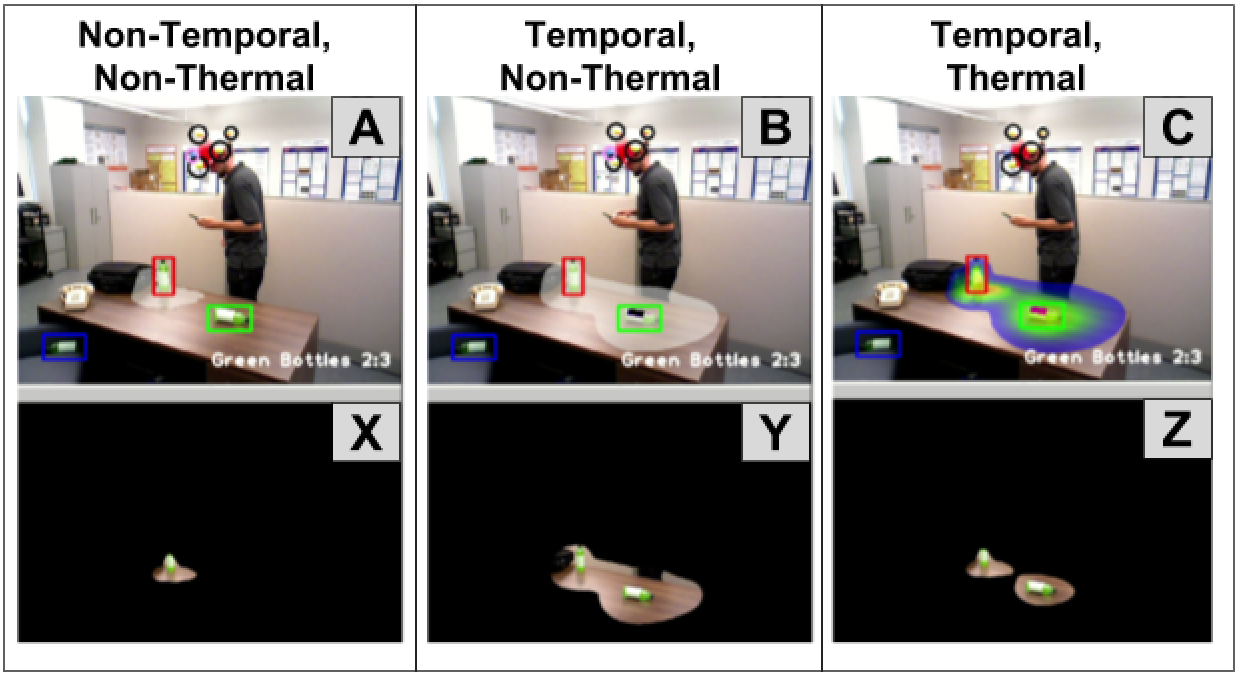


Figure 35: Comparing mapping techniques.

4.2.1 Temporal Thermal Map Advantages

The experiment in Section 4.2 helps highlight that the degree of saliency information provided by temporal thermal maps is superior to the non-thermal and non-temporal techniques. One flexibility of the thermal map implementation is that the extracted ROI image can be scoped dynamically at runtime for different levels of attention. What this means is different levels of the color gradient can be selected to strip out the ROI's saliency information. For example, the ROI image in Figure 35-Y represents 100% of the attentional map whereas the ROI in Figure 35-Z image represents a setting of 75% of the color gradient. Note that the blue ring of this thermal map represents the bottom 25% of the color gradient. This percentage can be adjusted at runtime, which allows the different levels of saliency in the ROI to be dynamic. If an external image classifier has trouble processing the ROI, this color gradient scope level can be adjusted to reduce the ROI size. This is possible because the thermal map is accumulated independently from the RGB image in this implementation of HPT4M.

4.3 Clustered Object of Interest Discrimination

The results in Section 4.2 highlight a key observation in this work which is that the temporal thermal mapping technique has a unique advantage over the non-temporal non-thermal methods for dynamically adjusting the map's saliency threshold level at run-time. This allows the system to be relatively agnostic about the subject's gaze envelope size parameter and allows for multiple OOI classification retries at different ROI threshold scoping levels without need for rebuilding the attention map. This key observation motivates a subsequent experiment to investigate how this advantage can be leveraged over the alternative approaches to attention modeling.

4.3.1 Clustered OOI Experiment Overview

For this experiment, a clustered OOI discrimination task is arranged such that the subject and Observing Agent are paired across from each other in a one-on-one table top work environment. This experiment environment emulates a shared team workspace and could serve as a precursor for collaborative tasks which could apply an attention model such as team assembly or learning by demonstration. In this experiment, the subject periodically observes a set of items and the HPT4M system is challenged to determine the subject's OOIs by using the gradient slicing technique to serve ROI images to an object classifier until the best scoring candidate OOI is found. Figure 36 shows the observing agent's point of view in the left frame and the subject's point of view in the right frame. There are six closely located objects of three different categories: tools, toys, and cups.

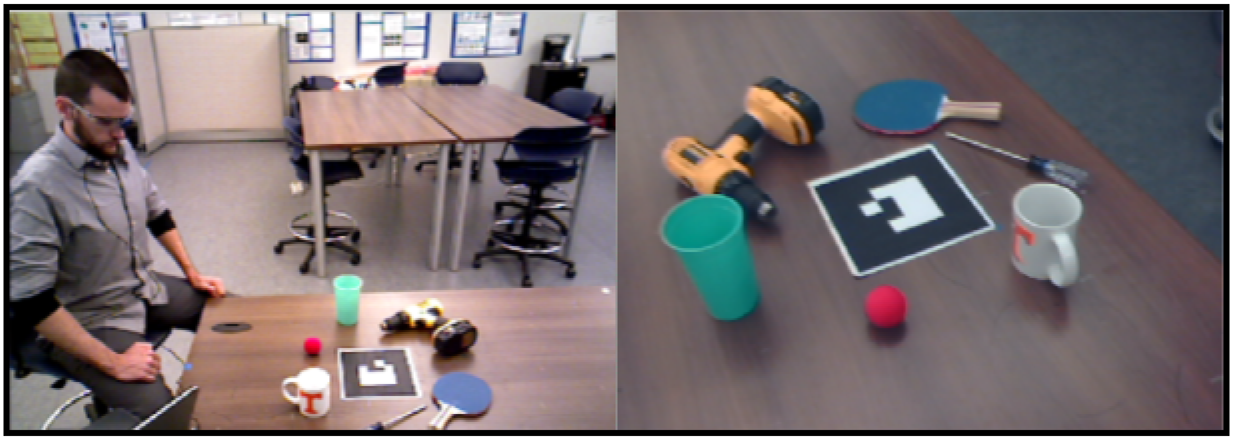


Figure 36: Clustered object of interest discrimination experiment.

4.3.2 Clustered OOI Alternative Head Pose Tracking

Since this particular experiment does not require aft perspective of the subject, it is possible to use an alternative tracking apparatus that allows for real-time verification of the subject's OOI and for the helmet marker apparatus to be removed. This also demonstrates that HPT4M is not dependent on any one particular head tracking approach. For this scenario, the user wears a pair of glasses with camera mounted in the center and the table is tagged with a fiducial marker. As long as both the subject and the OA keep the marker within view, both head poses can be localized within a common coordinate frame by using ARPose [Morris et al., 2010] to detect the location of the marker within each image frame. This allows for the OA to derive the subject's head pose within the RGB-D sensor frame and then render the attention map.

4.3.3 Clustered OOI Experiment Results

Table 1 shows the results for when the subject is currently observing the drill object, such as is shown in Figure 37. Each column shows the object classifiers scores for the set of objects, given the current thresholding level.

Table 1: Drill OOI classification scores for different map thresholds.

Current OOI=Drill	**Classification scores for gradient slicing percentages**				
Objects	**20%**	**40%**	**60%**	**80%**	**100%**
green cup	0.1192	0.0873	0.1221	0.3791	0.4245
utk cup	0.1361	0.1008	0.0152	0.0101	0.0598
paddle	0.0545	0.0259	0.023	0.0081	0.004
screw driver	0.0995	0.0224	0.0039	0.0049	0.0008
red ball	0.1689	0.1118	0.0969	0.0356	0.0133
drill	0.4219	0.6519	0.739	0.5622	0.4976

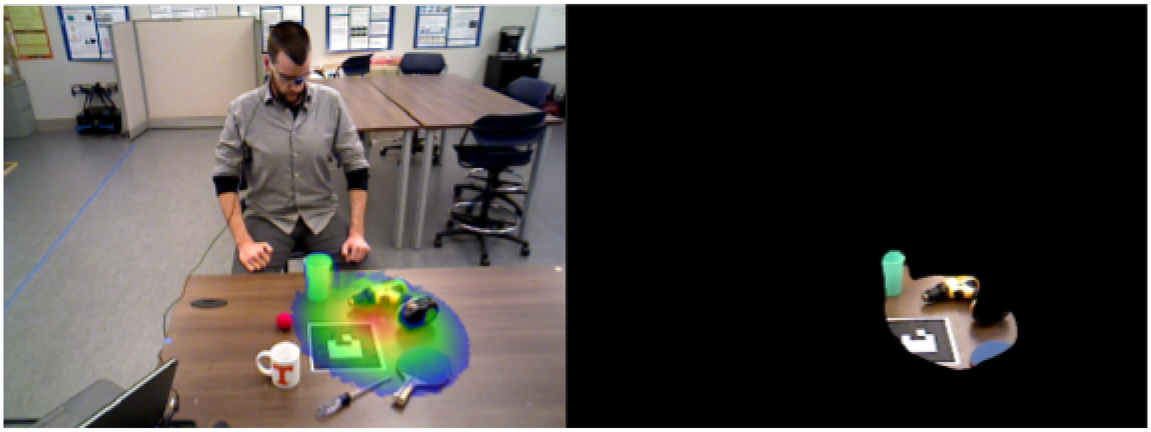


Figure 37: Clustered OOI temporal thermal map and an ROI with 80% threshold.

The 100% thresholding level uses the entire gaze envelope and makes it difficult for the system to discriminate the current OOI when the objects are closely clustered. This would be the result for the non-thermal mapping approach. As the gradient threshold is reduced, this essentially sifts the map and allows the classifier to extract a best scoring item. The thresholding level can essentially slice the map to account for the spacing between the clustered objects, thus it can dynamically adapt to the object's placement density in the environment. Objects that are more spaced out will begin to win the detection scoring with less thresholding. Closely clustered objects require tighter scoping until a clear winner is found.

The right image of Figure 37 shows a case where 80% thresholding still causes multiple objects to show up in the ROI. As seen in Table 1, reducing this level scopes the map more tightly around the drill and aids the classifier with a more accurate classification score. The 100% column shows confusion between the scores for the drill and green cup. The drill score becomes the clear winner with 80% or lower threshold but begins to degrade with 40% and below when too much information is lost by scoping the ROI too tightly.

4.3.4 Clustered OOI Experiment Discussion

The clustered OOI experiment has helped to demonstrate a clear advantage for temporal thermal mapping over alternative techniques. The gradient slicing of temporal thermal map regions of interest can improve the accuracy of an object classifier by reducing the scope of the ROI until a satisfactory result can be determined. It is important to note here that the subject's eye pose is assumed to match the head pose in this work. Further investigation into eye pose deviations from head pose is outside the scope of this work and is a candidate for future investigation.

4.4 Summary

This chapter has discussed the team search mission experiment, the clustered OOI discrimination experiment, and an evaluation of alternative mapping techniques. The search experiment shows that HPT4M is useful for implicitly discovering a human's OOIs, helping search for the OOIs, and labeling the OOI discovery states. The clustered OOI discrimination experient has showcased the thermal map gradient slicing technique's ability to improve an object classifier's accuracy of an ROI image for OOI determination. The successful proof of principles in the experiments, along with the validation testing in Chapter 3, support the posed hypotheses:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs.

Lastly, the experiment with alternatives techniques to temporal thermal mapping also reveals that for situations in team search missions where the subset of OOI types is already known, the extra computation for temporal thermal maps can be avoided by leveraging head pose trajectory tracking directly. Otherwise, temporal thermal maps provide a higher degree of scene saliency information than the non-thermal and non-temporal methods.

5 Implementation Challenges & Limitations

This thesis work encountered several critical challenges and hurdles that must be confronted when considering the use of HPT4M for real-world robotics application. These challenges include physics limitations of the depth camera sensor, scene object surface extraction from point cloud data streams, RGB-D camera calibration, and iris lighting dynamics.

5.1 Camera Physics Limitations

RGB-D cameras (see Section 3.2) have inherent physical limitations and thus constrain the use of HPT4M for a given sensor implementation. These sensors probe the scene depth by emitting an infrared (IR) light projection and timing the returned signal. This IR signal can become obscured by opaque surfaces and also ricochet off flat surfaces that have a high angle of incidence to the sensor. This causes an absence or obscurance of the return signal, which can effectively cause scene objects to be invisible or warped. Furthermore, the sensor is only designed for close range computer interfaces and is constrained to indoor use.

Figure 38 shows an output from from the RGB-D camera with the color raster on the left half and the IR depth raster on the right half. The black pixels in the depth frame represent null data due to lack of IR signal return. This capability limitation demonstrates that the sensor selected for the HPT4M application is critical. This issue can be avoided by careful scene crafting to avoid signal loss, which is sufficient for experimentation but not real-world applications.

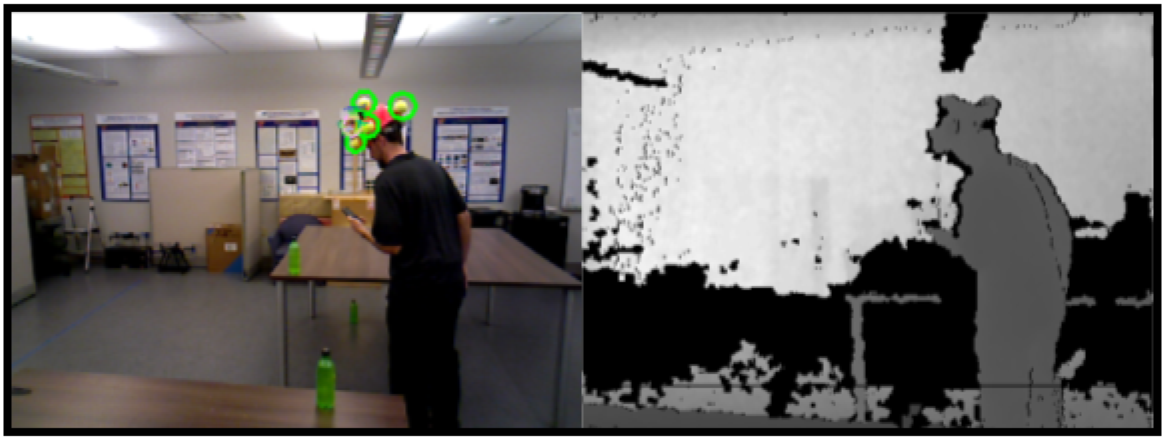


Figure 38: RGB-D camera color and depth frames.

5.2 Point Cloud X-Ray Phenomenon

The RGB-D camera sensor used in this thesis simply provides a raw point cloud data stream. The sensor does not naturally understand how the sampled points connect together to form scene surfaces or how to anticipate the shape of the unscanned backside of objects. The lack of surface knowledge and backside shape predictions result in the thermal map rendering upon the scene elements with a phenomenon similar to x-ray vision.

As can be seen in Figure 39 noted by the red star and arrows, the point cloud data scan results in an empty back face of the green soda bottle. This, in turn, causes a hole in the table, which inappropriately allows the thermal map under the table to be seen. Also note the orange star in Figure 39, which highlights an example of the x-ray phenomenon. This second lower blob should be blocked by the table's surface.

Clearly, a raw point cloud data set alone does not contain the geometric information necessary to correctly render the gaze projection in the scene and account for object surfaces and back sides. A valid approach to this problem would be two-fold. First, the point cloud data needs to be converted to a surface mesh. Secondly, the true knowledge of object back sides cannot be known until scanned. An approach for dealing with this could be to match recognized scene objects to a database of similar objects and build a virtual scene that contains full object sides and back faces. This application of a priori object knowledge could allow the correct thermal map to be rendered in the virtual scene. For situations where the scene contains novel items, the thermal map may be incorrectly rendered.

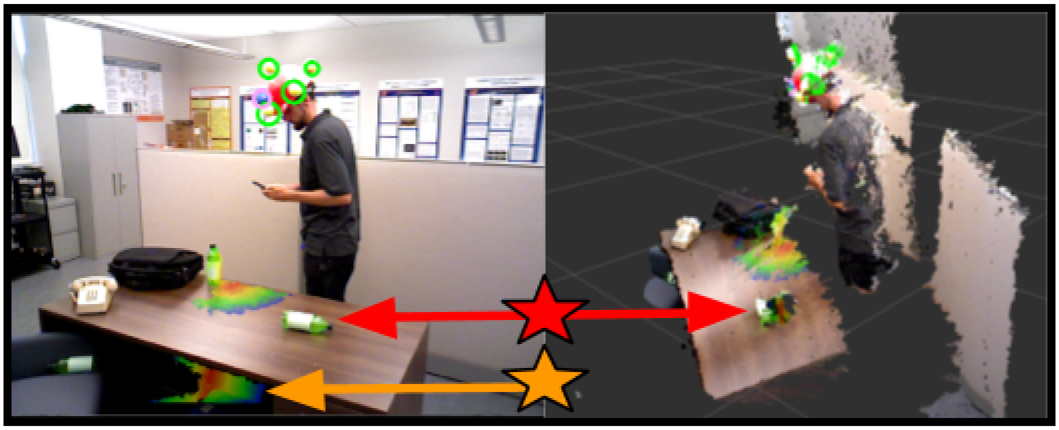


Figure 39: Point cloud scan causes object back face culling.

The simplifying solution used in this implementation of HPT4M is to detect the closest blob to the head coordinate and assume blobs further away are members of the x-ray phenomenon. As can be seen in Figure 40, the green star and arrow indicate the detection of the x-ray blob. These blobs are eliminated from the rendering and thus allow for the avoidance of the complexities related to point cloud stitching, point cloud-to-fragment conversion, and object a priori virtual scene rendering.

Again, such as was the case with the issue of the sensor physical limitations, scene crafting is also used to help mitigate the x-ray phenomenon issue in the experimental scene set up. The x-ray phenomenon only occurs in special cases where the observer sees multiple planes within the gaze direction of the subject. For example, in a fully convex or concave scene, the x-ray phenomenon would not occur.

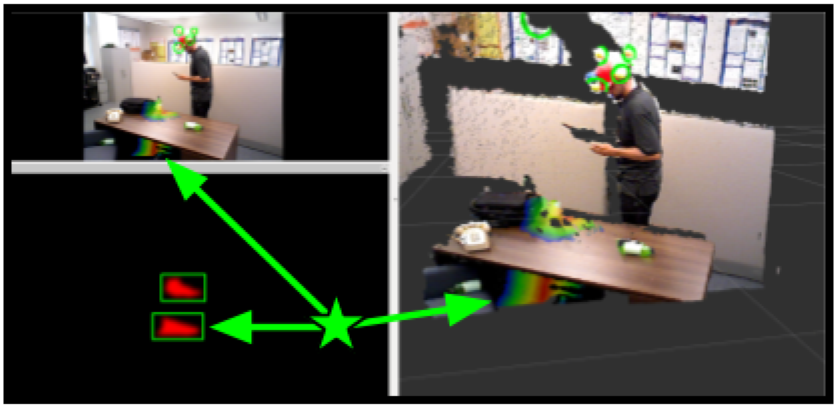


Figure 40: X-Ray blob detected for elimination.

5.3 Camera Calibration

It is critical that the RGB-D camera be properly calibrated such that the color and depth frames align properly in order to build an accurate point cloud. Figure 41 shows an example of a point cloud generated from a miscalibrated RGB-D sensor. The red star and arrows point to the tennis balls that are incorrectly mapped to the back of the room in the point cloud. As indicated by the red star arrows, this error causes the helmet tracking to be mislocated, which results in false head pose localization. This critical signal must be accurate in order for the thermal map projection to land correctly in the rendered scene. The solution to this issue is to properly calibrate the RGB-D camera and use the hardware depth registration driver settings.

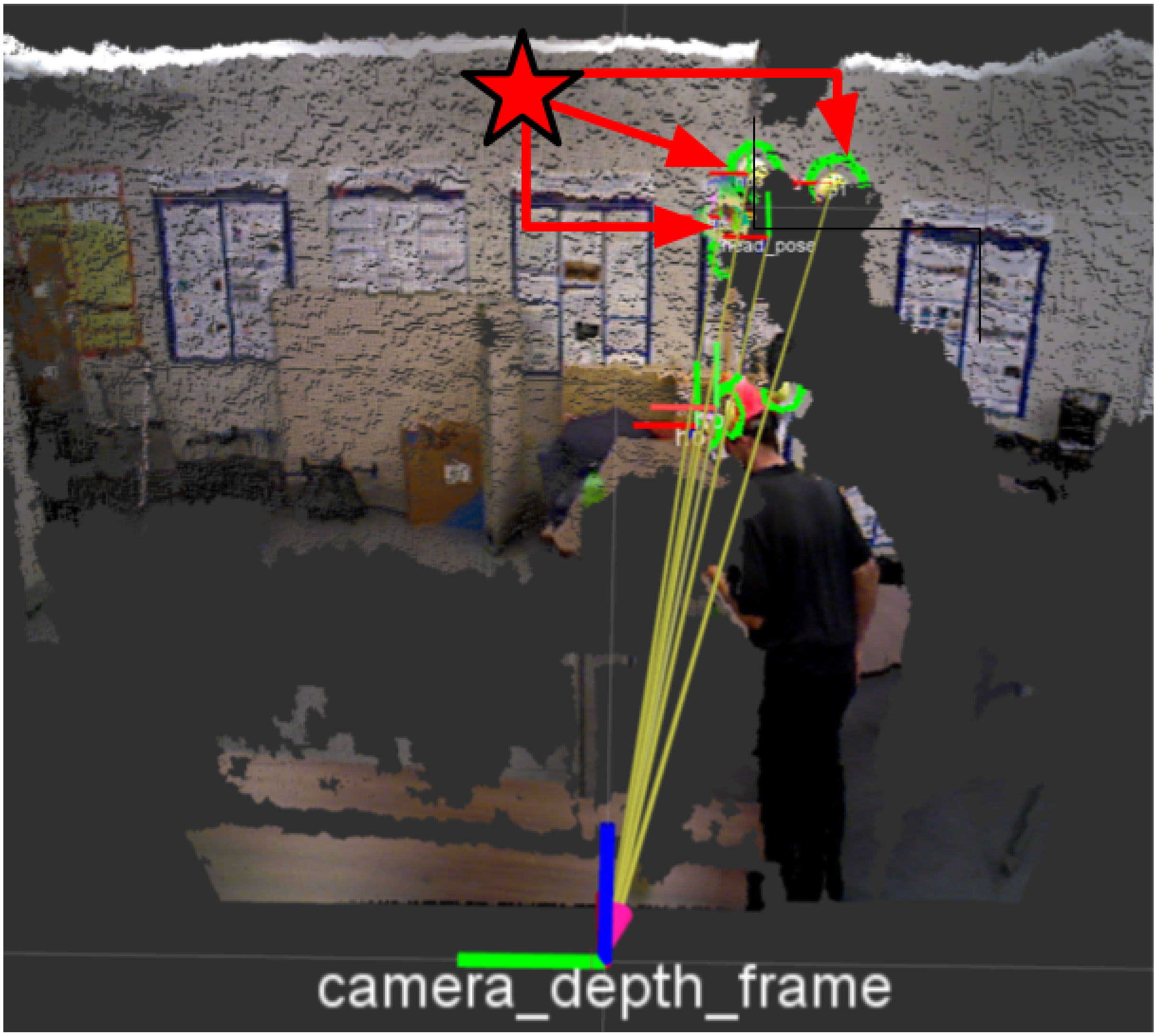


Figure 41: RGB-D camera miscalibration example.

5.4 Dynamic Camera Iris

The dynamic camera iris causes difficulty in tracking the yellow tennis balls. As the helmet changes position relative to the light ballast, the automatic camera iris causes the yellow color range to change significantly. Figure 42 highlights this issue with the red arrows in relation to the ceiling light ballasts.

The solution is to develop a dynamic runtime adaptation in the vision processing pipeline, which is discussed in further detail in Section 3.4.1. Essentially, this pipeline feature allows the thresholding for filtering the tennis ball locations to dynamically adjust as the scene parameters change over time with iris adjustments.

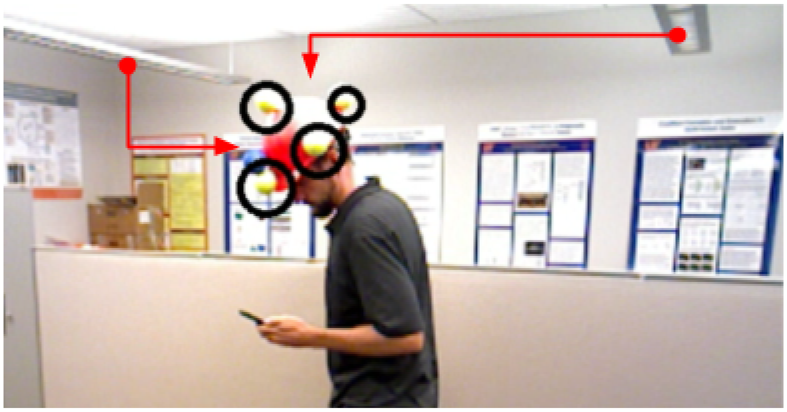


Figure 42: Dynamic camera iris in relation to the ceiling light ballasts.

5.5 Summary

Chapter 5 has discussed and shown some of the challenges and limitations encountered by the HPT4M implementation in this work. These challenges include physics limitations of the camera sensor, scene object surface extraction from point cloud data streams, RGB-D camera calibration, and iris lighting dynamics. Any real-world use of HPT4M must take these issues into account. Consequently, future advancements for HPT4M systems will likely need to apply surface extraction techniques for point clouds and backside fragment completion for partially scanned objects. Otherwise, the thermal map rendering will eventually encounter similar errors to the ones demonstrated in Section 5.2. Lastly, outdoor use of HPT4M is possible, but a different type of depth sensor will be required.

6 Conclusions

The goal for this thesis is to improve the capabilities of Human-Robot teams by making use of implicit communication. The included experiments conducted with the new head pose trajectory tracking tool demonstrate the capabilities of implicit peer attention modeling using temporal thermal maps. As shown for certain situations, HPT4M can be used by a robotic observing agent to:

- Implicitly discover a human subject's OOI type,
- Contribute to the search for OOIs, and
- Map the searched space.

Furthermore, it can be concluded that the proof of principles in the successful HPT4M experiment results support the posed hypotheses:

- Implicit information about a human subject's attention can be reliably extracted with software by tracking the subject's head pose trajectory, and
- Attention can be represented with a 3D temporal thermal map for implicitly determining a subject's OOIs.

This novel implementation of HPT4M for Human-Robot teams is directly inspired by the success of 2D monitor eye gaze tracking technologies. Broader impacts of HPT4M on Human-Robot teams could include:

- Reducing the amount of explicit commands that must be given by a human,
- Increasing the number of robots a human can manage at one time,
- Reducing the amount of time robot assets spend waiting for commands,
- Enabling new application domains for Human-Robot Teams, and
- Allowing human teammates to spend more time focusing on higher level mission objectives.

This thesis explores a small potential of HPT4M capabilities in Human-Robot team application and also shows some of the limits and hurdles involved with pushing the technology toward further real world applications. The experiment with alternatives techniques to temporal thermal mapping reveals that for situations where the subset of OOI types is already known, the extra computation of temporal thermal maps can be avoided by leveraging head pose trajectory tracking directly. In addition, it also reveals that HPT4M is able to provide a superior level of scene saliency information about the human subject's attention model. These proof of principles are foundational work for deeper investigation in future related research. This research could include human trials with the team search mission and other applications for implicit temporal thermal based attention modeling in Human-Robot teams, such as human activity prediction.

This thesis includes information on HPT4M implementation challenges and limitations with the intent that these discoveries will be valuable to those considering this technique for other applications. Despite the successes of HPT4M demonstrated in this thesis, several hurdles remain in order to be able to reliably make use of this technology in a real-world system. The physical sensor limitations of the RGB-D camera used in this work constrain the use to short range indoor applications with proper infrared light reflective surfaces. Also, in order to avoid the X-Ray phenomenon, HPT4M must be paired with a point cloud mesh algorithm and an object detection system that can render a scene based on prior object knowledge. Lastly, the bulky head pose helmet marker apparatus is not practical for real-world applications. It is recommended that any further investigation of this technology make strong considerations of these issues and limitations before pursuing HPT4M for use in an application.

In conclusion, a working HPT4M system is implemented, sample application experiments are successfully conducted with the new head pose trajectory tracking tool, discoveries of implementation challenges and limits are heeded, the posed hypotheses are supported, and foundational groundwork for future research has been laid. As shown by the proof of principles in this thesis, HPT4M is a valid technique for improving Human-Robot teams by facilitating implicit communication.

List of References

Abad M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.: “TensorFlow: Large-scale machine learning on heterogeneous systems”; Available from: tensorflow.org. 2015.

Benfold B, Reid ID: “Guiding visual surveillance by tracking human attention”, *The British Machine Vision Conference*, 10.5244, 2009.

Bradski G, Kaehler A: OpenCV [Internet]. “Dr Dobb’s Journal of Software Tools 2000”; Available from: http://mirror.sysu.edu.cn/wiki.ros.org/attachments/Events(2f)ICRA2010Tutorial/ICRA_2010_OpenCV_Tutorial.pdf.

Breazeal C, Kidd CD, Thomaz AL, Hoffman G, Berlin M: “Effects of nonverbal communication on efficiency and robustness in human-robot teamwork”, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp 708–713, 2005.

Broz F, Lehmann H, Mutlu B, Nakano Y: “Gaze in Human-Robot Communication”, *John Benjamins Publishing Company*, 2015.

Cao Z, Simon T, Wei S, Sheikh Y: “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, *CVPR*, 2017.

Cyberbotics: Webots [Internet]. cyberbotics.com 2017 [cited 2017 Jun 26]; Available from: https://www.cyberbotics.com/doc/guide/citing-webots?version=remidhum:reference_proto.

Duchowski A: “Eye Tracking Methodology: Theory and Practice”, *Springer Science & Business Media*, 2007.

Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ: “Automatic generation and detection of highly reliable fiducial markers under occlusion”, *Pattern Recognition, 4*7:2280–2292, 2014.

Google: Google Cloud Vision API [Internet]. google.com 2017 [cited 2017 Jun 21]; Available from: https://cloud.google.com/vision/.

Gredebäck G, Johnson S, von Hofsten C: “Eye Tracking in Infancy Research”, *Developmental Neuropsychology,* 35:1–19, 2009.

Microsoft: Microsoft Kinect [Internet]. Microsoft 2017 [cited 2017 Jun 21]; Available from: https://www.microsoft.com/en-us/store/d/kinect-sensor-for-xbox-one/91hq5578vksc/.

Morris B, Dryanovsk I, Dumonteil G: ARPose 2010 [cited 2017 Jun 26]; Available from: https://github.com/ar-tools/ar_tools.

Murphy-Chutorian E, Trivedi MM: “Head pose estimation in computer vision: a survey”, *IEEE transactions on pattern analysis and machine intelligence,* April 31:607–626, 2009.

NASA-JSC: Robonaut [Internet]. jsc.nasa.gov 2016 Oct 19 [cited 2017 Jun 21]; Available from: https://robonaut.jsc.nasa.gov/R2/.

Northrop Grumman International Inc.: Remotec [Internet]. northropgrumman.com 2013 Sep 2 [cited 2017 Jun 21]; Available from: www.northropgrummaninternational.com/wp-content/uploads/2013/09/Remotec-Andros-Titus.pdf.

OpenNI: OpenNI [Internet]. OpenNI 2017 [cited 2017 Jun 21]; Available from: https://structure.io/openni.

PackageInSight: Package InSight [Internet] 2016 Oct 18 [cited 2017 Jun 21]; Available from: http://www.packageinsight.com/.

Quigley M, Conley K, Gerkey B, Faust J, Foote T, Leibs J, et al.: "ROS: an open-source Robot Operating System", *ICRA workshop on open source software,* 2009.

Rethink Robotics: Baxter [Internet]. RethinkRobotics.com 2017 [cited 2017 Jun 21]; Available from: http://www.rethinkrobotics.com/baxter/.

Runxin M, Yang Y, Xiaomin Y: "Survey on Image Saliency Detection Methods", *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 10.1109:329-338, 2015.

Russakovsky O, Deng J, Su H, Krause J, Fei F, et al.: "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, 10.1007:211-252, 2009.

Sung J, Kanade T, Kim D: "Pose Robust Face Tracking by Combining Active Appearance Models and Cylinder Head Models", *Int J Comput Vis,* 80:260–274, 2008.

Szegedy C, Vanhoucke V, S Ioffe, Shlens J, and Wojna Z: "Rethinking the Inception Architecture for Computer Vision", CoRR, 1512.00567, 2015.

TobiiPro: Tobii Pro [Internet] 2017 [cited 2017 June 21]; Available from: https://www.tobiipro.com/.

Zanuttigh P, Marin G, Mutto CD, Dominio F, Minto L, Cortelazzo GM: "Time-of-Flight and Structured Light Depth Cameras", *Technology and Applications*, Springer, 2016.

Zhang H: "3D Robotic Sensing of People: Human Perception, Representation and Activity Recognition" [Internet] 2014 [cited 2017 Jul 12]; Available from: http://trace.tennessee.edu/utk_graddiss/2885/.

Vita

Corey Johnson was born near Huntsville, AL and grew up on a small farm North of Fayetteville, TN. During his undergraduate career, he participated in three NASA Robotics Academy Internships at Goddard Space Flight Center and also worked part time as a Student CNC Machinist. During school, he participated in four seasons of FIRST Robotics and four IEEE Hardware Challenges. He graduated from Tennessee Tech University in 2010 with a degree in Computer Engineering, a minor in Mathematics and an Outstanding Senior Design Team Award for the capstone project: *Solar Powered Autonomous Robot*. He began his professional career at Remotec and worked 4.5 years as an Embedded Software Engineer on Andros™ hazardous duty robots. Corey is currently working on a PhD in Computer Science at The University of Tennessee and is a member of the Distributed Intelligence Laboratory.