Real-time computer vision and deep learning for 3D environment modeling, camera network calibration and human-robot interaction using a robot dog

Zhigang Zhu[a,b], Jie Gong[c], Chong Di[c], Eltan Samoylov[a], Brandon Vasquez[a], Haiqiao Liu[c], Shengyuan Feng[c], and Fred Roberts[d]

[a]Department of Computer Science, The City College of New York / CUNY, New York, NY 10031, USA
[a]Department of Computer Science, The CUNY Graduate Center, New York, NY 10016, USA
[c]Department of Civil and Environmental Engineering, Rutgers University, Piscataway, NJ 08854, USA
[d]Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA

ABSTRACT

Video surveillance in public facilities, such as train and bus stations, airports, shopping malls, and sports arenas, is very important to public safety, both for identifying threats/terrorist attacks and implementing evacuation plans. The goal of this research is to explore the potential of using real-time computer vision and deep learning algorithms with a Boston Dynamics robotic dog Spot for the modeling of a large public venue, and in its collaboration and interaction with the 3D model of the large public venue, a network of surveillance cameras monitoring the area, and humans in the environment. The first step of our work is to explore and enhance the 3D vision navigation algorithms of the robotic dog to survey and map the area interactively, with real-time performance and georeferenced accuracy, so that the next time Spot comes to the area, it can automatically localize itself on the map. Then, to calibrate the network of cameras monitoring the area, no visual control points in the environment are needed; instead, by using the walking Spot to generate known 3D calibration target points accurate enough and widely distributed across the field of view of a camera, its intrinsic and extrinsic parameters can be estimated. Equipped with the 3D digital map with the calibrated surveillance cameras, the area can be automatically monitored through the collaboration of the camera network and Spot, which can be sent over to assist in understanding the situation as soon as a need is identified, by using real-time deep learning models such as YOLO models to detect humans in need and/or suspicious articles in the perimeter. Our work is closely related to the DHS mission with the ever-increasing security concerns in public venues, and has significant scientific and societal impacts, including enhancing research in digital twins, smart surveillance, robotics, and assistive technology.

Keywords: Digital twins, smart surveillance, robot navigation, camera calibration, assistive technology

1. INTRODUCTION

Video surveillance in public facilities, such as train and bus stations, airports, shopping malls, and sports arenas, is very important to public safety, both for identifying threats/terrorist attacks and implementing evacuation plans. Traditional video surveillance systems heavily rely on human operators to monitor activities, and therefore have many limitations. Such manual surveillance requires large amounts of tedious and time-consuming work and takes humans away from tasks that computers cannot solve. In addition, such work is prone to mistakes during long-term monitoring. It is also very challenging to identify unusual activities in a crowd scene by monitoring a bank of video screens.

Further author information: (Send correspondence to Z. Zhu)
Z. Zhu: E-mail: zzhu@ccny.cuny.edu, Telephone: 1 212 650 6248

The DHS Center of Excellence SENTRY (Soft-target Engineering to Neutralize the Threat RealitY) addresses the challenges of protecting soft targets and crowded places. The SENTRY vision to address threats to soft targets and crowded spaces is a suite of systems, which will function semi-autonomously with the capability to rapidly integrate and process data to provide real-time decision support to decision-makers (e.g., school principals) as they interact with first responders to detect, deter and mitigate targeted violence. Aligning with SENTRY's vision, the CCNY DHS Summer Research Team (SRT) has worked with its host institution Rutgers, a partner in SENTRY, to study the integration of data from stationary sensors (security and surveillance cameras) installed in the infrastructure with mobile sensors (including cameras) on persons or robots.

Many researchers in both developing and studying surveillance systems have focused on automating video surveillance and human and anomaly identification.[1–3] On the other hand, studies have been performed in the social sciences in identifying vulnerable populations such as old adults, children, and people with disabilities.[4,5] Recent progress in AI and robotics has enabled a lot of interesting applications with great social impact. In particular, the integration of quadruped robots with large language models such as ChatGPT provides a lot of potential, in guided tours,[6,7] clearing out explosives or inspecting radioactive environments,[8] brain-computer interaction,[9] assisting the visually impaired.[10,11]

However, there are still a number of unmet challenges. First, the majority of the cameras in large public facilities are not calibrated, and it would be a daunting task to manually align them with the 3D models of a scene when hundreds of cameras are installed. Second, using mobile robots as a way to calibrate cameras as well as assist people in need is a very interesting topic that has not been explored. Third, the integration of technology and social science studies in identifying vulnerable populations can provide a convergence solution to the problem that is highly relevant to the DHS mission. We envision that the research and development would also be extended to screening of suspicious individuals and behaviors through the analysis of clothing, actions, eye movements and interaction patterns with other people and the robot.[12,13]

2. OVERVIEW OF THE WORK

The CCNY-Rutgers SRT research focused on exploration of the potentials of Boston Dynamics Robotic Dog - Spot for the modeling of a large public venue such as a stadium and a museum, and in its collaboration and interaction with the 3D model of the large public venue, a network of surveillance cameras monitoring the area, and humans in the environment. The work includes three closely related and intertwined topics (Figure 1):

1. 3D mapping and digital twin creation for navigation at scale with a robotic dog as the modeling agent.
2. Calibration of a network of surveillance cameras with the robotic dog as the calibration target.
3. Intelligent assistance of people in need with the robotic dog.

The overall idea is to use the robotic dog Spot to survey and map the area interactively, so that the next time that Spot comes to the area, it can automatically localize itself with the map. Then to calibrate the cameras monitoring the area, no visual control points in the environment are needed; instead, by using the walking Spot to generate known 3D calibration target points and observe them in the field view of a camera, the camera's parameters can be estimated. Equipped with the digital map with known camera information, the area can be automatically monitored through the collaboration of the camera network and Spot, and as soon as a need is identified, Spot can be sent over to assist the situation. In this way, the video and audio feed can be processed and coded in the frontend, and privacy issues can be lessened.

In all the three steps, we aim to achieve realtime implementation, by leveraging the available onboard computational and sensory resources of the robot dog spot for 3D mapping (Section 3), the lightweight computational requirements of Unity3D and calibration algorithms on a laptop or desktop (Section 4), and the computing and sensing capacity of a mobile device held by a user for the multimodal interfaces, as well as the cloud based services of LLMs (ChatGPT) and the lightweight deep learning models such as YOLO models for object detection and content analysis (Section 5).

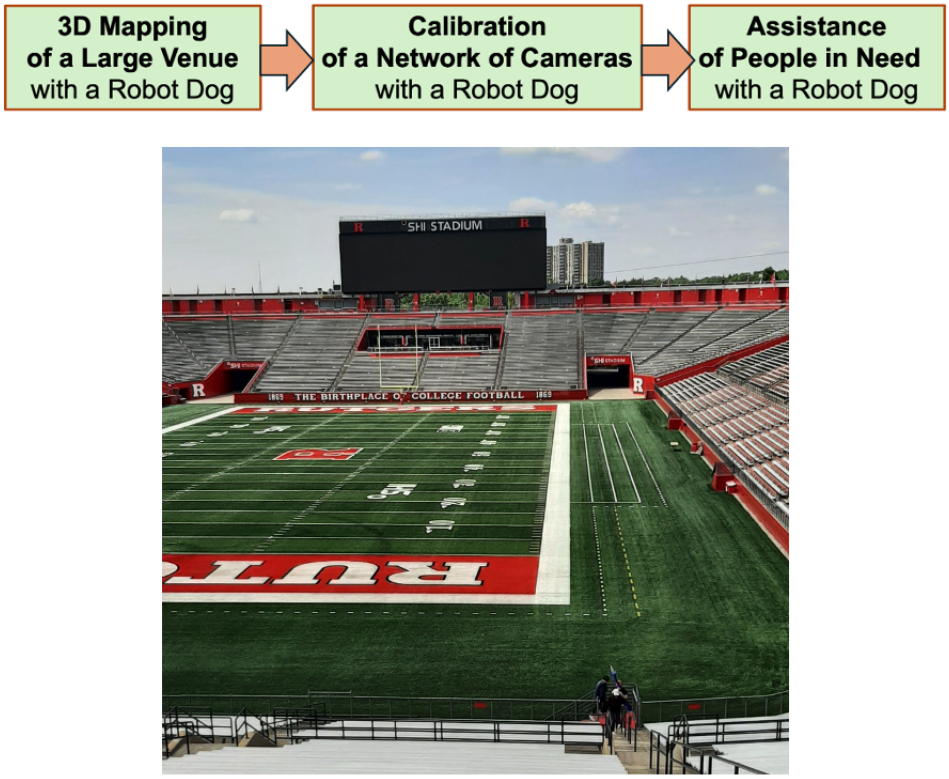

Figure 1. Three building blocks of the proposed framework: mapping of a large public venue, calibration of a camera network, and assistance for people in need, all with a robotic dog. Bottom: a scene of the Rutgers' SHI Stadium - fields, stages and surveillance cameras.

3. 3D DIGITAL TWIN CREATION FOR NAVIGATION AT SCALE

A large body of literature over decades has contributed to the knowledge of 3D modeling of a large 3D space for navigation and localization. For an indoor environment, vision-based methods are the more prominent approaches, especially for robotic applications, Simultaneous Localization And Mapping (SLAM) is the standard of the field. In the past, the collaborative team of CUNY and Rutgers has explored methods in using static LiDAR to create an accurate digital twin model of an indoor environment,[14] or simply a smartphone's visual sensors to create a lightweight 3D model,[15] for assistive navigation of people who are blind or have low vision. In this research, since we hope to unleash the potentials of the robotic dog Spot for multiple purposes as listed above, we have explored the capacity and limitation of the mapping function provided for Spot, namely the Autowalk,[16] and identified research issues for building 3D maps for navigation at scale.

3.1 Autowalk Fundamentals

During Autowalk, Spot intelligently navigates along a preset route to capture data or perform useful actions on a site.[16] To help Spot succeed, we place Spot-recognizable fiducials (printed placards similar to QR codes with the same standard established by the APRIL Robotics Laboratory of University of Michigan[17,18]) around the site and learn how to build robust, repeatable Autowalk missions. An Autowalk mission consists of two stages. The first stage is intended to model the as-built environment (known as map recording), which is primarily done via manual control (and via automatic exploration in our future studies). During the operation of the modeling stage, Spot localizes itself and perceives its surroundings by integrating pictures from its onboard perspective cameras and the 3D point cloud data live streamed by a Velodyne LiDAR sensor (which is mounted on Spot's "lower back" as shown in Figure 2). Maps are created with two components: a point cloud dataset for the as-built environment and a topological graphs illustrating the locations and connections between the Spot waypoints

Figure 2. Spot on the move.

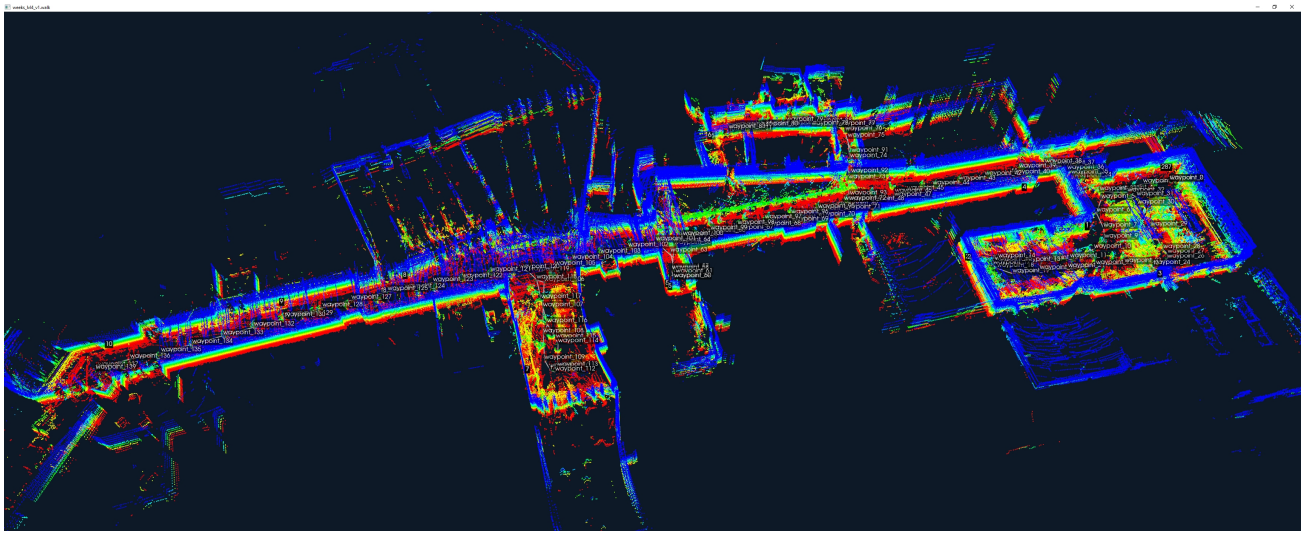

Figure 3. A map created by Spot's Autowalk with a graph of waypoints as the map representation and attached by point cloud data.

(Figure 3). The second stage, known as the navigation stage, can start at any time once a map is recorded (after the modeling stage). During this stage, Spot can autonomously determine the most efficient route and travel to any waypoints described by the map. There are inevitably going to be a great number of waypoints while Spot is modeling large as-built environments, so Spot's Autowalk controling program allows the waypoints to be automatically generated and placed as the following:

- At 2 m intervals along straight paths;

- When either event occurs within a 0.3 m path segment or Spot turns more than 30 degrees;
- Elevation of Spot changes more than 0.3 m; and
- A user-specified Action (together with the robot location or pose) is recorded.

In practice, in the manual tablet control mode of an Autowalk mission, a user initializes the map recording by starting Spot at a location with a fiducial in its sight and controls the robot to the next point-of-interest for waypoint establishment. The user can name each waypoint with particular meanings, such as a room number, a turn on the corridor, or under a staircase. While the map recording stage is about to complete, Spot does not need to close the loop of its trajectory so that the Autowalk mission can end at an arbitrary location.

Then in the navigation stage, the user can specify a destination and Spot will plan a path with pre-selected waypoints to the destination and automatically go to the destination following the prebuilt route map. Spot will follow all the waypoints along its path, but it can be adaptive to small changes of the environment between two waypoints (such as moving tables or the existence of static or dynamic obstacles) as long as a path can be found automatically between them. During mission replay (navigation), Spot calculates its position by comparing features in its current sensor data with the features in the data snapshots taken at each waypoint during mission recording.

3.2 Autowalk Limitations

In our early experiments, we have found several limitations.

(1) Fiducials are specially designed images, similar to QR codes, that Spot uses to localize itself at the initial position as well as to align its internal map with the world around it. Fiducials are generally used to mark specific point-of-interest, for instance, docking stations, and are required at the beginning of any Autowalk mission.

(2) The original navigation algorithm needs to see all waypoints along a planned route. Spot automatically compensates for deviations in its path and small changes in the environment, but large discrepancies may require Operator intervention.

(3) Actions can be added to Autowalk missions during the Autowalk recording process by selecting "+Add Action" on the controller. Spot only supports up to 100 total Actions per mission, which are also pre-defined.

(4) For automatic mapping, fiducials need to be placed around the site for Spot to explore the area. The models may not be metric accurate since it is a relative model.

3.3 Research Topics Identified and Tested

We proposed to integrate building information model (BIM), Static LiDAR Scan and the Spot Autowalk map for generating and updating the digital twin of a large venue for survey, navigation and assistance. We note that a BIM is created in the design stage. A static LiDAR model is accurate but static. Spot Autowalk gives a reliable and updated model but the accuracy is relative for the purpose of the robotic navigation. Here are the number of things we identified and partially tested.

(1) The accuracy of the Spot localization and the improvement for mapping in scale. We measure Autowalk's accuracy whenever it matches a waypoint. For this purpose, we set up multiple fiducials on the walls that both our static LiDAR scanner and Spot can see and measure. Then by comparing the accurate LiDAR scan and the Autowalk map for the 3D locations of the fiducials, we can obtain the accuracy of Spot localization. Our experiment shows that in exploring an indoor area, the localization accuracy of Spot depends on the distances it travels: for an Autowalk mission of a total traveling distance of 175 meters (574 feet), the error in its localization from 50 meters to 175 meters (164 to 574 feet) increases from 5 cm to 40 cm (2 to 16 in) on average (Figure 4).

Based on our past experience,[15] we proposed the integration of two novel ideas. The first one is to divide the large space into overlapping regions, and create an Autowalk mission for each region, and connect the maps of all the missions into a large global map. In this way, the creation of digital twins can be implemented for large-scale

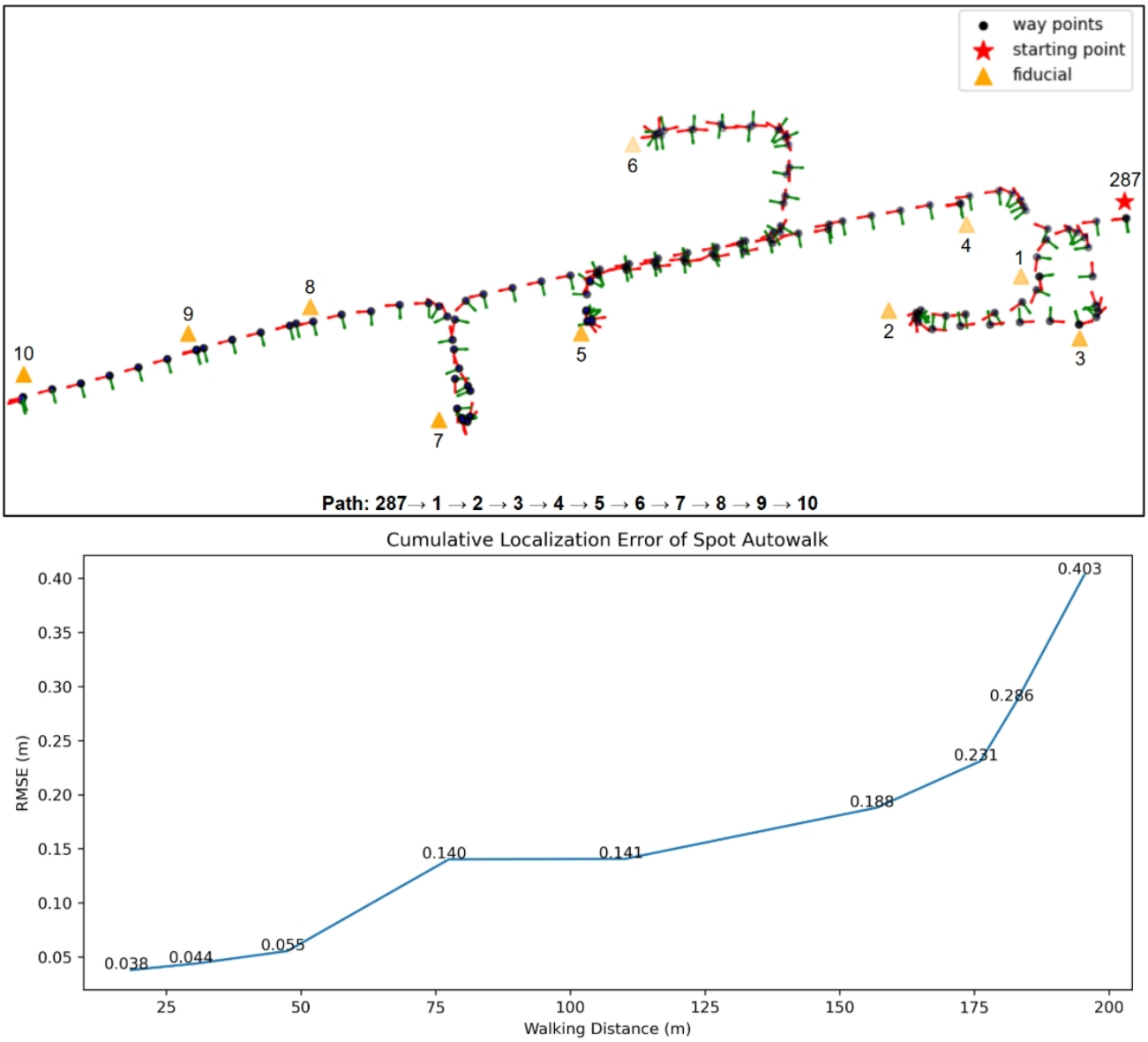

Figure 4. The trajectory and localization accuracy of during the Spot Autowalk. Top: Fiducials (represented by triangles) whose ground-truth locations are known from the accurate LiDAR scan in Figure 5), plotted on the Autowalk map with a graph of waypoints (the robot center is represented by a black dot; red and green arrows are showing the Spot front and left respectively). Bottom: The graph illustrates the cumulative localization error (SLAM) for the Spot Autowalk, which is evaluated against the ground truth fiducial locations.

environments. The second idea is to use an accurate floor plan (e.g. from BIM), or the digital twin created by the static LIDAR scan (Figure 5), allowing us to interactively or even automatically match the local Autowalk Spot locations and its point cloud with the accurate floor plan, so that the two maps can be aligned. If Spot walks in an outdoor environment, the GPS locations can also be used as the global 3D localization measures for model alignment.

(2) Using Spot for surveying accessible features. Thanks to Spot's capability to walk in all-terrain in 3D spaces, the localization and mapping information can provide rich information to identify ramps, stairs, elevators

Figure 5. Alignment of Spot's Autowalk map with the static LiDAR scanned model. The texture-mapped point cloud, known as the world point cloud or ground truth point cloud, is collected by our high-definition static laser scanner. Red dots denote the Spot positions recorded (by the robot) during the Autowalk mission. The point cloud data captured by two separate systems - the static LiDAR system and the Spot sensor system - are aligned based on the fiducials.

of a site, using both its 3D model and its images. We proposed to expand the Spot action set by automatically adding new actions related to our task, namely accessibility for people in need, including those who have visual, hearing, mobility and mental challenges by detecting important landmarks in the scene using deep learning models.

(3) Reducing or even removing the use of fiducials. The current support of the Spot Autowalk needs to place a fiducial to start a mission. And every time a playback is implemented, Spot has to stay close to the fiducial. We would like to first remove the need of Spot going back to any fiducial to start a playback, by enabling Spot to recognize a waypoint and localize itself and be able to go to any predefined destination that is created during the mapping stage. Then we hoped to prepare Spot for the spatial knowledge generated from our accurate static LiDAR model so that Spot can not only rapidly localize itself at the start of its mapping mission, but also maintain a high self-localization accuracy by aligning its local Autowalk model with the static LiDAR model during the mission.

4. CALIBRATION OF A NETWORK OF SURVEILLANCE CAMERAS WITH SPOT

4.1 Calibration Basics

Detecting static or moving objects (humans, vehicles, suspicious articles) in a public venue with the views of surveillance cameras does not need camera calibration if only a detection or recognition is needed. However, a calibration is necessary if we want to get more accurate information of the detected target, such as its location, its distance and its size in 3D space. Given a set of known 3D-2D matches, both the intrinsic parameters (the focal length, the center of the image, the aspect ratio, etc.) and the extrinsic parameters (i.e. the pose including position and orientation, represented by a rotation matrix and a translation vector) of a camera can be estimated.

4.2 Current Limitations

Camera calibration has well-developed algorithms in many computer vision packages, such as OpenCV and Matlab. However, there are three practical issues in using them in a large venue such as a stadium with many cameras. First, an algorithm typically needs a well-designed calibration target such as a checkerboard to provide known 3D control points (Figure 6) presented in the view of the camera. This also means that human operators

have to go to the field to set up the calibration target. Second, the calibration thus done can only provide the estimation of the intrinsic parameters of the camera, such as the center of the images, the focal length, and the aspect ratio. The camera pose estimation with extrinsic parameters - the 6 degrees of freedom (DOF) rotation and translation parameters - will be relative to the checkerboard and it is another challenge to relate the checkerboard with a global coordinate system. Finally, it would be best that the control points for camera calibration cover the field of view (FOV) so the camera parameters would be useful for measurements of locations and sizes in the camera's view. However, making a checkerboard with a large size and from a large distance to the camera is impractical.

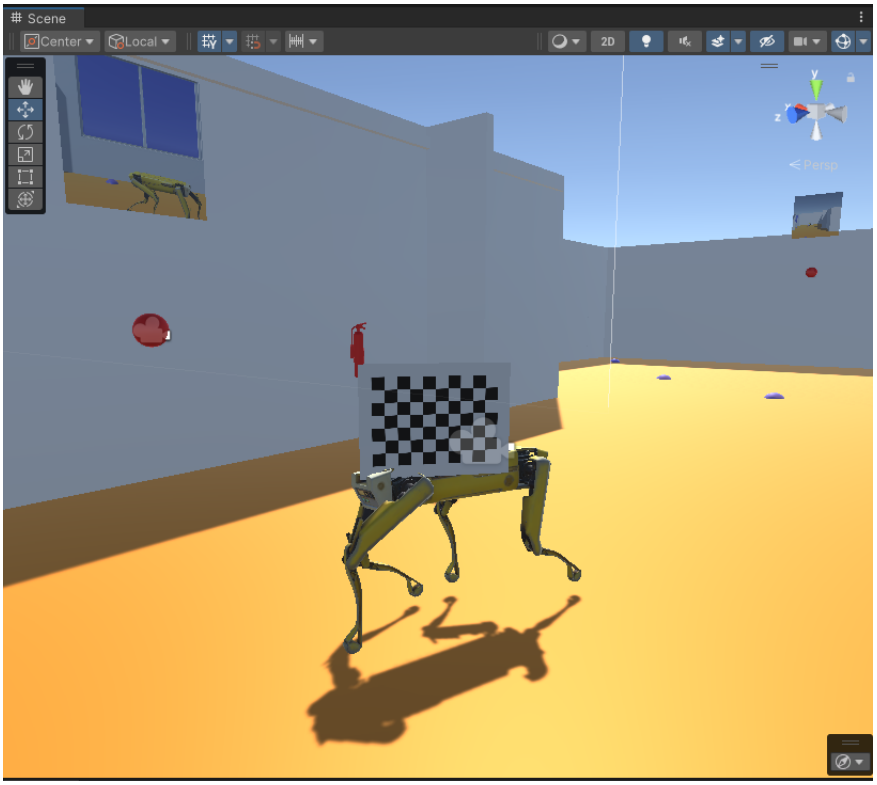

Figure 6. A checkerboard carried by Spot for camera calibration.

4.3 Research Topics Identified and Tested

We proposed to use the robotic dog Spot as the calibration target. Imagine that an operator sitting in front of a bank of screens for the surveillance cameras can remotely send the robot to the field, and under the FOV of each camera. Then the robot can either autonomously walk around the area without human intervention or be interactively controlled to send it to ensure its locations are distributed across the FOV of the camera. Whenever spot is automatically detected in an image of the camera, its 3D location can be obtained from the Autowalk system, which is also in the global world coordinate system. Then a set of 2D-3D matches are automatically obtained, and with more than 6 pairs of matches, the full set of the camera intrinsic and extrinsic parameters can be estimated.

For achieving this goal, we identify the following research tasks.

(1). Realistic simulation in Unity3D. For developing and evaluating the accuracy and robustness of camera calibration using a walking Spot as the calibration target, we first developed a Unity3D virtual environment with realistic scenes,[19] the robot dog, and virtual surveillance camera. In order to streamline the pipeline from virtual

to real, we have proposed to develop digital twins of real environments such as a stadium, museum and a research facility. Figure 7 shows a virtualized environment (aka digital twin) of the Rutgers Sustainable Infrastructure Laboratory where experiments of mobile robots (drones and walking robots) are carried out. In the environment, the virtual space includes a 3D virtual environment matching the real measurement and setups of the lab, the robot dog Spot, and three virtual surveillance cameras (Figure 8).

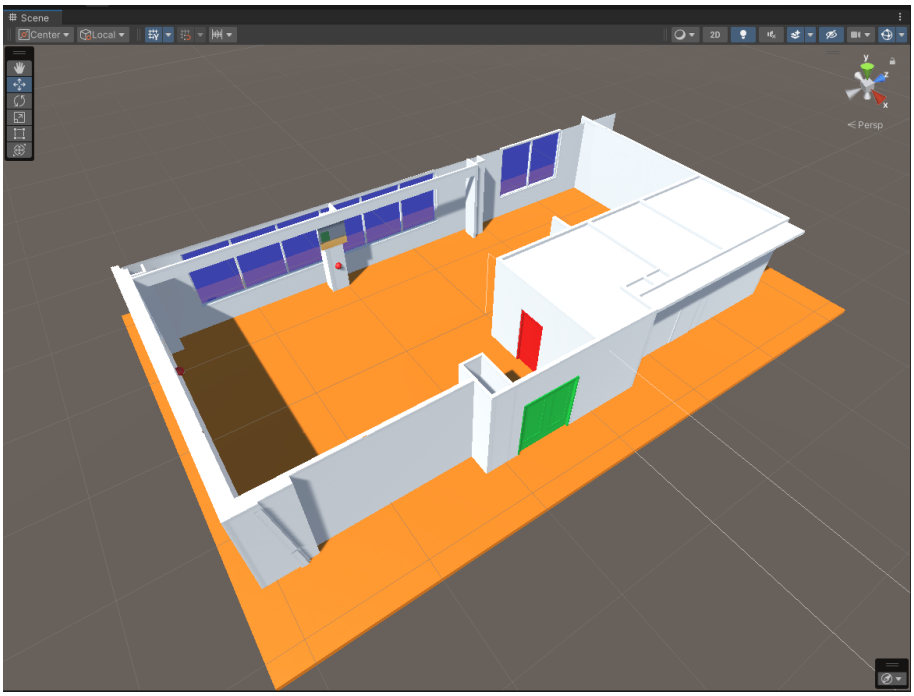

Figure 7. A virtualized environment of an infrastructure lab.

(2). Camera calibration in the virtualized environment. Camera calibration in this work is to find the relations of the three coordinate systems: the world, the robot and each of the surveillance camera, as well as each camera's specification including its focal length (one parameter), the center of the image (two parameters), and its aspect ratio (one parameter). While the relation between the world coordinate system and the robot coordinate system has been pre-calibrated when the robot Spot mapping the environment (Figure 9), we only need to find the relation between the camera coordinate system and the world coordinate system, represented by a 3x3 rotation matrix and a 3-dimensional translational vector.

We look into three calibration approaches and evaluate their performance in various situations: Spot with a calibration checkerboard, Spot as one point in motion, Spot with multiple key points. We evaluated what are the limitations of each approach, in terms of the field of view (FOV), 3D and 2D measurement errors, and the requirements in feature extraction of the calibration targets in images.

A. Spot with a calibration checkerboard. When the robot Spot carries a planar checkerboard (Figure 6) with known relations of the co-planar 3D points on the checkerboard in the robot coordinate system since they can be estimated accurately using the static Lidar scanning, we can use the robust algorithm initially developed by.[20] This only requires more than two images be taken when the checkerboard is with different orientations. This can be achieved when Spot walks around within the FOV of a surveillance camera, by intentionally changing its orientations of the planar checkerboard, whose parameters can be estimated by Spot's Autowalk related to the world coordinate system. Then the calibration algorithm will provide both the intrinsic parameters and the extrinsic parameters of each planar checkerboard's orientation. Hence, the camera's extrinsic parameters with respect to the world coordinate system can be calculated with the known orientations of the moving checkerboard in the world. This will be an especially effective method when the robot can be close to the camera.

Figure 8. Spot walking in the VE with surveillance camera views. Top: the developer's view of the space. Bottom: a view from a camera on top of the robot Spot which includes the front part of the robot (yellow), and a surveillance camera represented by a red dot on the back wall. Three insets show three views of the scene from three surveillance cameras, and the robot dog is in the view of the first camera (left).

B. Spot as one point in motion. if the FOV of the camera is large, and the robot is far away from the camera, for example in the stadium (Figure 1), the robot as seen in each image might be very small, and either a checkerboard or the feature points on Spot may not be very distinguishable from a far distance. In this case, we proposed to detect Spot as a whole (for example with color segmentation since the body of the robot is yellow), and use the centroid of the detection as a feature point. When it moves in a large field of view like the SHI Stadium, due to its walking and stair-climbing capacity, the robot can generate 3D feature points in the three-dimensional space, on the field and on stages, so a general calibration algorithm as in OpenCV can be used.[21]

C. Spot with multiple key points or different height configurations. When Spot is walking in close range of the camera to be calibrated, it can be seen as a 3D moving object, even though it moves on a 2D surface. As an example, we can use the features of the robot dog, both on the top of the body, and the legs touching the floor (Figure 9). In this way, the calibration method eliminates the need of any unnatural calibration target except the robot dog itself. As another example, since SPOT can change its postures with different heights linearly, we can generate 3D points that are not one a single plane even though it walks on a flat surface. Figure 10 shows the results of a simulation where Spot can generate 3D-2D matches when walking in 3D space, even when a single key point is extracted from every SPOT posture/location.

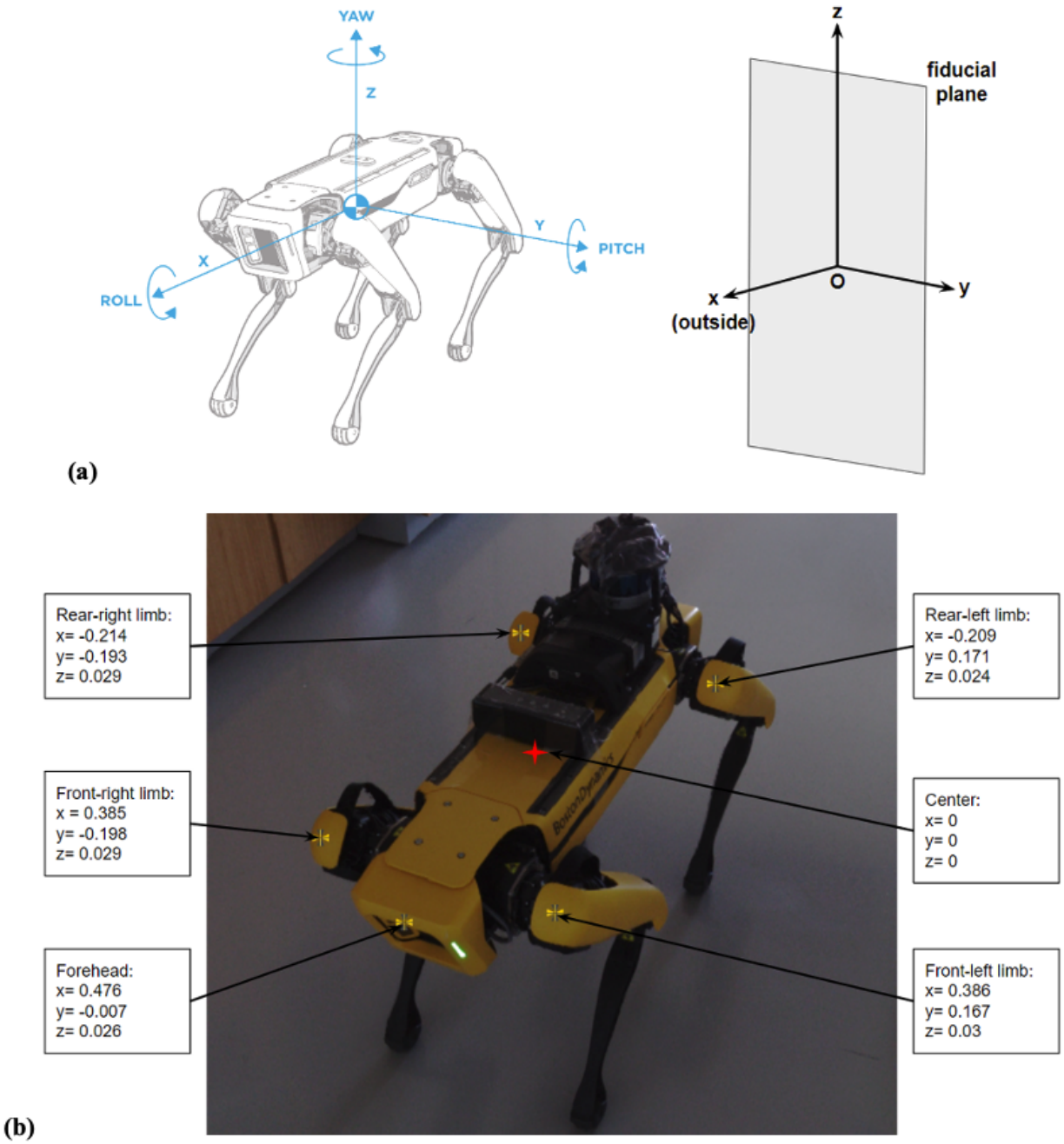

Figure 9. The robot and world coordinate systems in (a) and the key points of the robot Spot in (b), as measured (in meters) in the world coordinate system established by a fiducial on the wall.

(3). Camera Calibration in the Real World. The final step is to apply the calibration algorithms for the surveillance cameras in a real-world environment. For facilitating the data collection and interactive calibration process, we would like to visualize the real robot dog Spot and the camera feeds of the surveillance system in the digital twin of the real environment of interest. Since the robot's pose can be estimated automatically, we know exactly when it is in it's walking mode, and when it comes in the FOV of a surveillance camera. Then the operator can remotely control the robot to walk in a desired pattern, while the 2D images of the robot are

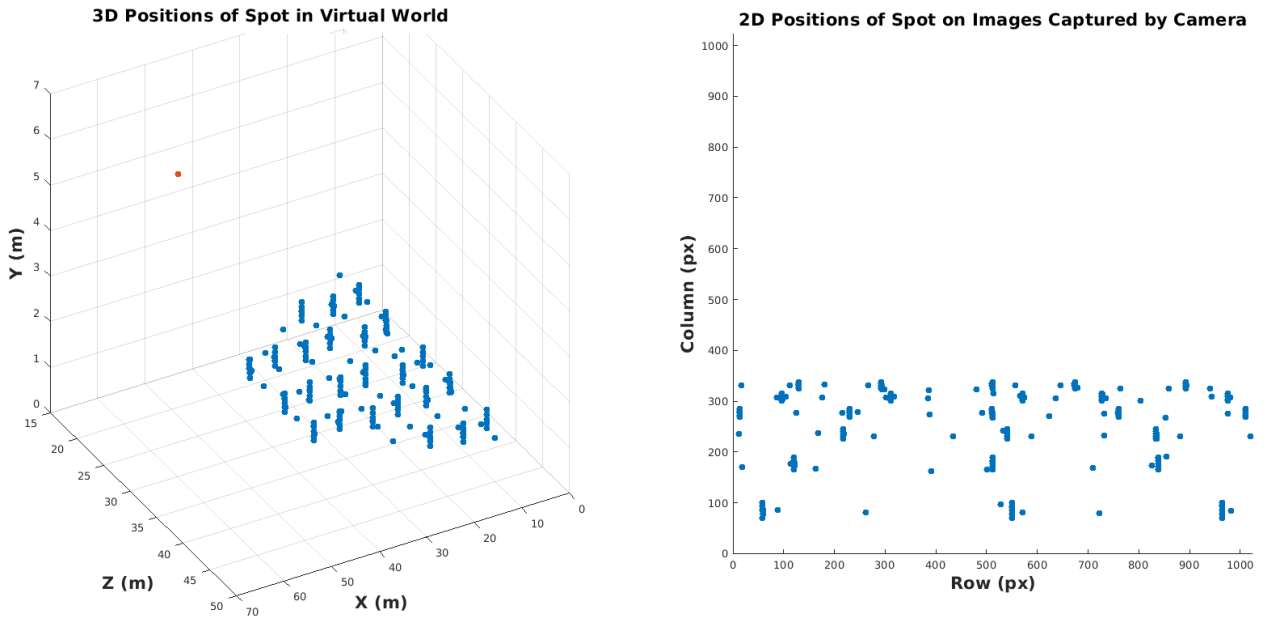

Figure 10. Camera calibration using the walking Spot changing its heights when walking on a flat surface. Left: the 3D control points collected when the virtual SPOT is walking on the floor but changing its heights. Right: the corresponding 2D points as seen in the images of a size of 1024×1024 pixels.

being detected on the images of the surveillance camera. After obtaining 3D-2D matches of Spot in motion, the camera parameters can be estimated via calibration and the video feed can be better projected in the virtual 3D space, the same as the case in our full VE simulation (Figure 8). This will be future work on our follow-on project.

5. INTERACTION OF ROBOT SPOT WITH PEOPLE AND ENVIRONMENTS

5.1 Research Topics in Interactive Interaction

With the technical preparation of mapping and calibration as two key milestones, the intelligent interaction of the robot dog with people and the environment can be achieved. This includes three major components:

(1) Detect and localize humans from surveillance cameras. Since the environment has been modeled as a 3D digital twin with its surveillance cameras calibrated, detected humans and any unusual activities can be localized and tracked across multiple cameras, using deep learning models such as YOLO for detection. We proposed to develop algorithms to identify and re-identify people in need or in suspicion and share the information with operators to take action.[22–24]

(2) Send Spot over for assistance. The operators might send Spot over for assistance in a situation. Given the position of the event, Spot can automatically navigate itself to the destination without human intervention, and with the cooperation of the surveillance cameras, which also monitor the robot, Spot can find the individual who needs help and intervention.

(3) Interaction of Spot and Human via speech and with ChatGPT. The final step is the interaction between Spot and the individual who needs assistance. We are developing a speech interface for the robot to talk via speech-to-speech service with the individual, allowing for a conversational approach to the interaction. This interaction will be made possible by feeding user queries to ChatGPT[25–29] for relevant and human-like

responses, such as how to navigate a stadium or information on an art piece in a museum. This consists of the following three tasks (detailed below): (T.1) The system allows a user to interact with the Spot robot with multimodal channels via an iPhone or Android mobile device (Figure 11). (T.2) We customize ChatGPT for specific purposes by fine-tuning the text-generation model in a particular environment. (T.3) The system allows for the input of images, audio, and text as prompts using the GPT 4 model in collaboration with how Spot could respond to its surroundings and vocalize it.

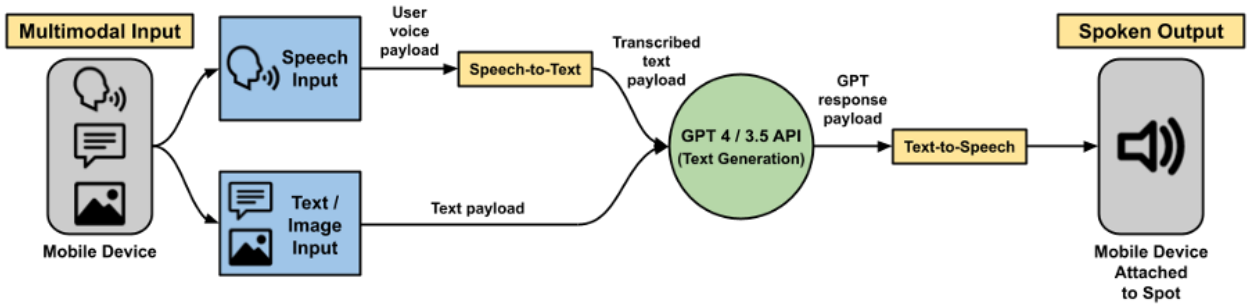

Figure 11. Multimodal interface for the robot Spot and the large language model ChatGPT.

5.2 Research Tasks in ChatGPT Interaction

Task T.1. User multimodal interface. The goal of Task T.1 is to allow a user to interact with the Spot robot with multimodal input (text, speech, or image) through a mobile device such as an iPhone or Android device (Figure 11). This will be accomplished by two interfaces. (i) Using the Python and JavaScript programming languages we will create a web interface to communicate with the OpenAI API. (ii) The user will be able to simulate interactions with the quadruped robot via a mobile device interface where multimodal input will be processed through the text-to-speech pipeline and vocalized via a speaker–or mobile device–mounted on the Spot robot, similar to the dual iPhone interface in.[9] (iii) The API that will be employed to achieve these tasks is the ChatGPT 4 model for text generation[25] given text and/or image input, speech recognition Whisper model,[26] and text-to-voice model.[27] All of these will be provided by the OpenAI platform, streamlining and facilitating compatibility across the various models being used.

Task T.2. Customization of ChatGPT for tasks and users. In order to customize ChatGPT for specific tasks such as security alerts, assistance during an evacuation, or for user(s) with vulnerabilities, we would like to fine-tune the ChatGPT model for the tasks and users in their respective environments (museum, transportation center, stadium, etc.). This will require that we investigate data collection, data labeling, and model training for the customized ChatGPT.[28] For example, suppose an individual who is blind or has low vision (BLV) visits a museum. In that case, the information from the articles as well as the models of the museum will be fed into the ChatGPT model so that the robotic dog Spot can answer specific questions about the exhibit articles or security information for the user regarding accessibility zones, safety exits, etc. In addition, it will be able to tailor the answer in an appropriate format, style, and language for the BLV user, such as voice feedback and vibration reminders. The system will also allow the user to replay the answer in a user-friendly manner.

Task T.3 Read images as input prompts. In the Follow-Up work we plan to do, we envision that the system will allow for the input of images as prompts, e.g. using the GPT 4 model in collaboration with the results of image analysis for Spot to respond to its surroundings and vocalize it using the same pipeline as Task T.1. While image input for GPT 4 is new and not as sophisticated in its responses as text inputs,[29] it will still be worthwhile to explore how GPT 4 could process and make sense of the images captured by Spot to assist in threat detection or aiding vulnerable populations.

5.3 Real-time Object Detection with YOLO

Here we show some results in detecting and localizing humans and navigation signages in public venues from surveillance cameras and/or the onboard cameras of the robotic dog SPOT. Since each scene has been modeled as a 3D digital twin with its surveillance cameras and SPOT's onboard cameras are calibrated, detected humans and objects and any unusual activities can be localized and tracked across multiple cameras. Figure 12 shows a few examples of real-time detection of people, obstacles such as chairs, suspicious items such as a suitcase, and other objects, using a deep learning model YOLO v11.[30] In this figure, we show detection results from both the left and right cameras of SPOT. In the left-camera image, a chair (with its boundary and label in purple) and a person (with the boundary and label in blue) are detected and extracted from the image. In the right camera image, a suitcase (with its boundary and label in green) is detected. The speed of the detection is from 110 ms to 190 ms per frame when executed on a computer with a 13-th gen Intel(R) Core(TM) i7-1360P processor.

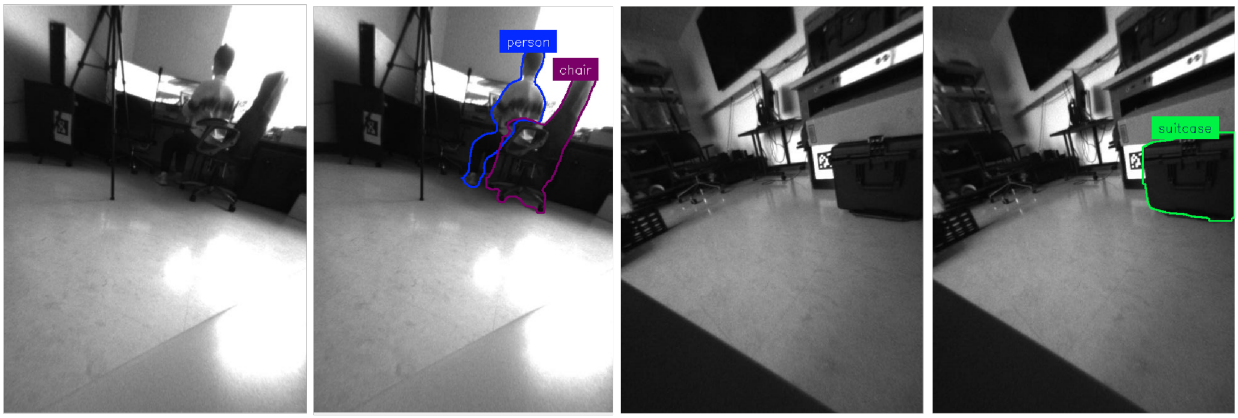

Figure 12. Real-time object detection on images from SPOT's left and right cameras using a YOLO v11 model (image from the left camera, detection results for the left camera image, image from the right camera and the detection results).

6. CONCLUDING REMARKS

The goal of this research is to explore the potential of the Boston Dynamics robotic dog called Spot for the modeling of a large public venue such as a stadium and a museum, and in its collaboration and interaction with the 3D model of the large public venue, a network of surveillance cameras monitoring the area, and humans in the environment. The overall idea is to use the robotic dog Spot to survey and map the area interactively, so that the next time Spot comes to the area, it can automatically localize itself with the map. Then to calibrate the cameras monitoring the area, no visual control points in the environment are needed; instead, by using the walking Spot to generate known 3D calibration target points and observe them in the view of a camera, the camera's parameters can be estimated. Equipped with the digital map with the calibrated network of surveillance cameras, the area can be automatically monitored through the collaboration of the camera network and Spot, and the results of the monitoring can be sent over to assist in understanding the situation as soon as a need is identified. The research can support the protection of soft targets and crowded places with aerial/ground agents.

ACKNOWLEDGMENTS

This article was developed under an appointment to the DHS Summer Research Team Program for Minority Serving Institutions, administered for the U.S. Department of Homeland Security (DHS) by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between DHS and the U.S. Department of Energy (DOE). ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-SC0014664. This document has not been formally reviewed by DHS. The work of Gong, Di, Feng, Liu and Roberts has also been supported by the U.S. Department of Homeland Security under Grant Award Number 22STESE00001-04-00 from DHS Office of University Programs via Northeastern University. The

views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DHS, DOE, or ORAU/ORISE. DHS, DOE and ORAU/ORISE do not endorse any products or commercial services mentioned in this publication. The work of Zhu, Samoylov and Vasquez has also been supported by the National Science Foundation (NSF) through Awards #2131186 (CISE-MSI) and #1827505 (PFI), and Disability and Accessibility Services at the CUNY Central Office of Student Affairs.

REFERENCES

[1] Socha, R. and Kogut, B., "Urban video surveillance as a tool to improve security in public spaces," *Sustainability* **12**(15), 6210 (2020). https://doi.org/10.3390/su12156210.

[2] Wu, Y., Zhao, J., Yu, N., and Feng, R., "Indoor surveillance video based feature recognition for pedestrian dead reckoning," *Expert Systems with Applications* **173**, 114653 (2021). https://doi.org/10.1016/j.eswa.2021.114653.

[3] Zhang, P., Tao, Z., Yang, W., Chen, M., Ding, S., Liu, X., Yang, R., and Zhang, H., "Unveiling personnel movement in a larger indoor area with a non-overlapping multi-camera system." arXiv (2021).

[4] Wingate, M. S., Perry, E. C., Campbell, P. H., David, P., and Weist, E. M., "Identifying and protecting vulnerable populations in public health emergencies: addressing gaps in education and training," *Public Health Rep.* **122**(3), 422–6 (2007). https://doi.org/10.1177/003335490712200319.

[5] Kuran, C. H. A., Morsut, C., Kruke, B. I., Krüger, M., Segnestam, L., Orru, K., Nævestad, T. O., Airola, M., Keränen, J., Gabel, F., Hansson, S., and Torpan, S., "Vulnerability and vulnerable groups from an intersectionality perspective," *International Journal of Disaster Risk Reduction* **50**, 101826 (2020). https://doi.org/10.1016/j.ijdrr.2020.101826.

[6] Roth, E., "Boston dynamics turned its robot dog into a talking tour guide with chatgpt," *The Verge* (2023). https://www.theverge.com/2023/10/26/23933213/boston-dynamics-robot-dog-spot-top-hat.

[7] Wessling, B., "Boston dynamics turns spot into a tour guide with chatgpt," (2023). https://www.therobotreport.com/boston-dynamics-turns-spot-into-a-tour-guide-with-chatgpt/.

[8] Petkauskas, V., "Chatgpt injected into boston dynamics' spot," *Cybernews* (2023). https://www.theverge.com/2023/10/26/23933213/boston-dynamics-robot-dog-spot-top-hat.

[9] Kosmyna, N., Hauptmann, E., and Hmaidan, Y., "A brain-controlled quadruped robot: A proof-of-concept demonstration," *Sensors* **24**(1), 80 (2024). https://doi.org/10.3390/s24010080.

[10] Chen, Y., Xu, Z., Jian, Z., Tang, G., Yangli, Y., Xiao, A., Wang, X., and Liang, B., "Quadruped guidance robot for the visually impaired: A comfort-based approach," in [*2023 IEEE International Conference on Robotics and Automation (ICRA)*], 12078–12084 (2023). doi: 10.1109/ICRA48891.2023.10160854.

[11] Sivacoumare, A., S, S. M., Satheesh, S., Athul, T., V, M., and Vinopraba, T., "Ai quadruped robot assistant for the visually impaired," 1–5 (10 2021).

[12] CORDIS, "Roborder: autonomous swarm of heterogeneous robots for border surveillance." Research and Innovation Community Platform of the European Commission. Last update: 12 April 2022.

[13] Contreras, R., "Robo dogs and ai inspectors might be coming to the border," *Axios* (2023). https://www.axios.com/2023/12/12/border-patrol-us-mexico-ai-robo-dogs-drones-inspectors.

[14] Gong, J., Feeley, C., Tang, H., Olmschenk, G., Nair, V., Yu, Y., Zhou, Z., Yamamoto, K., and Zhu, Z., "Building smart transportation hubs with internet of things to improve services to people with special needs," in [*Transportation Research Board (TRB) 96th Annual Meeting*], (2017).

[15] Zhu, Z., Chen, J., Zhang, L., Chang, Y., Franklin, T., Tang, H., and Ruci, A., "iassist: An iphone-based multimedia information system for indoor assistive navigation," *International Journal of Multimedia Data Engineering and Management (IJMDEM)* **11**(4) (2020).

[16] Autowalk, "Getting started with autowalk." https://support.bostondynamics.com/s/article/Getting-Started-with-Autowalk. (Accessed: 06/20/2024).

[17] BostonDynamics, "About fiducials - boston dynamics spot sdk." https://support.bostondynamics.com/s/article/About-Fiducials-77114. (Accessed: Jan 29, 2025).

[18] APRILRoboticsLaboratory, "Apriltag - april robotics lab - university of michigan." https://april.eecs.umich.edu/software/apriltag. (Accessed: Jan 29, 2025).

[19] Unity3D, "Unity3d, real-time development platform." https://unity.com/. (Accessed: Feb 03, 2024).

[20] Zhang, Z., "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11), 1330–1334 (2020). doi: 10.1109/34.888718.

[21] OpenCV, "Opencv, the world's biggest computer vision library." https://opencv.org/. (Accessed: Feb 02, 2024).

[22] Khan, S. M., Javed, O., Rasheed, Z., and Shah, M., "Human tracking in multiple cameras," in [*Proceedings Eighth IEEE International Conference on Computer Vision*], (2001). https://doi.org/10.1109/ICCV.2001.937537.

[23] Idrees, H., Shah, M., and Surette, R., "Enhancing camera surveillance using computer vision: a research note." arXiv (2018). https://doi.org/10.48550/arXiv.1808.03998.

[24] Iguernaissi, R., Merad, D., Aziz, K., and Drap, P., "People tracking in multi-camera systems: a review," *Multimedia Tools and Applications* **78**, 10773–10793 (2019).

[25] OpenAI, "Text generation models. openai developer platform." https://platform.openai.com/docs/guides/text-generation. (Accessed: Feb 02, 2024).

[26] OpenAI, "Speech to text. openai developer platform." https://platform.openai.com/docs/guides/speech-to-text. (Accessed: Feb 02, 2024).

[27] OpenAI, "Text to speech. openai developer platform." https://platform.openai.com/docs/guides/text-to-speech. (Accessed: Feb 02, 2024).

[28] OpenAI, "Fine-tuning models. openai developer platform." https://platform.openai.com/docs/guides/fine-tuning. (Accessed: Feb 02, 2024).

[29] OpenAI, "Chatgpt-4 model. openai developer platform." https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo. (Accessed: Feb 02, 2024).

[30] Khanam, R. and Hussain, M., "YOLOv11: An Overview of the Key Architectural Enhancements," (Oct. 2024). arXiv:2410.17725.