Vision and Language Navigation Using Minimal Voice Instructions

Ansh Shah
Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai,India
anshshah2111@gmail.com

Parth Kansara
Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai,India
parthkansara@gmail.com

Parth Meswani
Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai,India
parthmeswani29@gmail.com

Prof. Prachi Tawde
Professor, Information Technology
Dwarkadas J. Sanghvi College of Engineering
Mumbai,India
prachi.tawde@djsce.edu.in

Abstract— The proposed system aims to design an algorithm that can be used to navigate any 3-D mapped environment, using the Matterport 3D Simulator by giving only minimal voice instructions. During the training phase, the nodes of a selected environment are traversed sequentially in the Simulator and an object recognition algorithm is applied on the panorama at each node. This helps in identifying and tagging the objects in the vicinity of each viewpoint. For the testing phase, a natural language instruction, specifying the goal location is taken as input. The goal location is identified from among the various viewpoints in the 3D environment by matching it to the tags generated in the testing phase. A shortest path algorithm is employed to navigate from the starting location to the goal location. The proposed system focuses on the implementation of the algorithm which combines natural language processing and computer vision and can be employed by agents for indoor navigation.

Keywords— Indoor Navigation, Computer Vision, Natural Language Processing, Matterport 3D

I. INTRODUCTION

Navigating in an indoor environment is a complex task because of the accuracy that it demands and the inability of the outdoor technologies to deliver such accuracy. Since GPS has been rendered inaccurate through walls and roofs, alternate technologies for indoor navigation have been a topic of research for many years. While many have advocated the use of sensor-based approaches like SLAM [5,6], others have tried to integrate computer vision models into navigation algorithms [1,2,7]. The Matterport 3D Simulator [2] provides a large-scale reinforcement learning environment based on real imagery. Most recent advancements have integrated natural language instructions, along with computer vision techniques for navigating within the indoor environment. Anderson et al. [2] presents the Room-to-Room (R2R) dataset for visually grounded natural language navigation, which however includes detailed elaborate instructions.

However, a realistic model should be able to understand and navigate using minimal instructions, such as “Go to the sofa” as opposed to long complex navigation instructions [1,2] such as “Go straight down the hallway, turn left into the bedroom and wait near the sofa”. The proposed system will focus on reducing the length as well as the complexity of the sentence, which can be as short as just the goal location.

In the training phase, the proposed algorithm traverses through the nodes in the navigation graph generated by the Matterport 3D simulator [2]. At each node, the Matterport 3D dataset [9] contains a panoramic image. The proposed system applies YOLO [8] on each node’s panorama to identify and tag objects visible from that viewpoint. These tags are stored in the memory and are used to identify the goal location during the testing phase.

The input natural language instruction is first converted to text using the Google speech-to-text API. Following this, a LSTM is used to extract the goal location from the text. Once the goal location is identified, the proposed system matches it with the tags generated in the training phase to identify the goal node in the navigation graph.

After this, the improved Dijkstra's Shortest Path Algorithm [10] is applied to find the path to be followed from the starting node to the goal node.

DOI: 10.5281/zenodo.5090533
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

II. LITERATURE REVIEW

Visual Question Answering is one of the initial systems which has attempted to design a multi-discipline AI system, by combining Natural Language Processing and Computer Vision. It describes a model that can be used to answer complex questions pertaining to an image. Answering these questions requires the model to develop an understanding of the background details and the underlying context of the image. The task can be interpreted as a visually grounded sequence-to-sequence translation problem and has served as crucial groundwork for future research. This study has laid out the foundation for developing navigation algorithms, which overlap machine vision and natural language processing.

A. Vision and Language Navigation

The system works with the Matterport 3D dataset, which is the largest currently available RGB-D dataset. It has the most extensive depth image collection which includes multiple navigable points and allows multiple trajectories for simulating motion. Each navigable viewpoint has a panoramic image which captures the entire sphere of the visible scene from that viewpoint, except the poles. Each panorama is constructed from 18 RGB-D images, captured at the height of an average person. It is a diverse dataset, including scans from houses, offices, restaurants and shops among others.

The simulator constructed allows an agent to navigate through any of the scans available in the Matterport 3D dataset. It allowed to move between different viewpoints and adopt a pose, defined in terms of the following parameters:

1. 3D position, which is the 3D position of the viewpoint where the agent is present
2. Heading, which ranges from [0, 360) degrees
3. Elevation, which ranges from [-90, 90] degrees

On each timestep t, the simulator generates an RGB image, which corresponds to what the agent observes at that particular timestep.

To ensure that the agent only chooses navigable viewpoints, and obeys the physical constraints like walls and floors, the simulator defines a set of possible viewpoints that can be reached in the next step, at each timestep t. For this, the simulator uses a weighted undirected graph over panoramic viewpoints, G = {V, E} where the edges correspond to allowed navigation between the viewpoints. The weights of the edges represent the straight-line distance between the two viewpoints. It also eliminates any edges over 5m in length, to ensure the motion is localized.

This paper also introduces the R2R task and dataset. The R2R task provides the agent with a natural language instruction which is to be followed for navigating from the source node to the goal node in the Matterport 3D simulator. Movement at each timestep is based on the image observations of the agent, resulting from the movement at the previous timestep. Based on this task, 21,567 navigation instructions were collected from volunteers which makes up the R2R Dataset.

B. Chasing Ghosts: Instruction Following as Bayesian State Tracking

This system extends the Matterport3D Simulator to support depth image outputs. Additionally, the natural language instruction is interpreted as a sequence of expected actions and observations for the agent. Based on this intuition, the task of locating the goal location is formulated as Bayesian state tracking. It demonstrates credible results on the VLN task and uses an approach that depends less on the navigation constraints.

The proposed instruction-following agent has a mapper, a filter and a policy. The mapper builds a semantic spatial map of the surrounding environment. The filter decides the most probable trajectories and the goal locations in the spatial map. The policy performs a chain of actions to reach the predicted goal location.

The filter is the key contribution of this paper, which formulates the instruction following as a Bayesian state tracking problem. Given the starting state, the semantic spatial map developed by the mapper, and a chain of latent actions and observations, an end-to-end differentiable histogram filter is trained to predict the most probable trajectory taken by a human. The sequence of observations and actions required is extracted from the input natural language instruction, using a sequence-to-sequence model with attention mechanism.

For predicting the goal location in VLN, the system outperforms a strong LingUNet baseline. On the full VLN task, it has a success rate of 32.7%, which is a credible result. The highlight of this system is that apart from the policy, the entire pipeline is independent of the simulator and does not require nav-graphs. This brings it closer to real world implementation, which does not include nav-graphs.

Once the goal location is predicted, the set of actions to reach the goal location is to be predicted. VLN[2] models an Long Short Term Memory-based sequence-to-sequence architecture with an attention mechanism which predicts a probability distribution over the next set of possible actions. Chasing Ghosts[1] predicts a probability distribution on the action space using a two-layered neural network. LingUNet[x] generates an action using recurrent neural networks.

C. Chasing Ghosts: Instruction Following as Bayesian State Tracking

Peter X. Liu et al [3] talks about Sphinx-4, an offline Java speech recognition system which consists of FrontEnd, a Linguist , a Decoder and a Configuration Manager. Hidden Markov Models[3] are used for feature extraction and acoustic models are constructed on these features for language modelling. Separate speaker and follower models [4] have proven to be more efficient where the instruction speaker model is used in training as well as testing phase. The speaker model is used to synthesize new instructions and implement pragmatic reasoning. Experiments show that all three components of this approach—speaker-driven data augmentation, pragmatic reasoning and panoramic action space—dramatically improve the performance of a baseline instruction follower which more than doubles the current benchmark success rate.

Table 1. Comparative study of Computer Vision

Characteristics	**VQA- Visual Question Answering**	**Vision-and-Language Navigation: Interpreting visually grounded navigation instructions**	**Chasing Ghosts: Instruction Following as Bayesian State Tracking**
Dataset	MS COCO	Matterport3D	Matterport3D
Methodology	Model develops an understanding of the background as well as the context of the image and attempts to answer questions based on visual aspects of the image	Interpretation of natural language instructions based on the agent's vision is modelled as a sequence-to-sequence translation problem	Formulates the question of finding the target location as a Bayesian State tracking problem
Algorithms	Sequence to sequence neural network	Sequence to sequence neural network	Bayes filter

Table 2. Comparative study of NLP

Characteristics	**Speech Recognition Engine and Robotic Control Unit [3]**	**Speaker- follower model for Vision and Language Navigation [4]**
Dataset	Binary representation of speech files.	limited number of route pairs and navigation instructions.
Methodology	The user can direct the robot by either talking directly or talking into a mic, and the command is extracted and fed to the robot via wireless network connection. [3]	The paper uses a speaker model to (1) generate new instructions for data augmentation and to (2) implement pragmatic reasoning, which can estimate how accurately the candidate's chain of actions can elucidate an instruction. [4]
Algorithms	HMMs for predictive modeling. They allow us to predict a sequence of unknown (hidden) variables from a set of observed variables.	Student-forcing reinforcement learning algorithm for training.

The current systems have made benchmark progress with the novel approaches employed to solve targeted problems. Majority of them have employed a probabilistic model to predict the goal location and the set of actions for it.

One of prime observations is that the systems provide only a discretized action space, which is not accurate in a real-world scenario.

More so, the systems have been majorly confined to the existing Matterport 3D dataset and the Matterport3D simulator. This is because of the dependency of the systems on RGB-D images, which are essentially RGB images with a depth component.

Also, the majority of systems requires a detailed instruction containing a set of actions and observations.

But intuition follows that an efficient system would be able to function on shorter instructions, as minimal as only containing the goal location.

III. Proposed system architecture

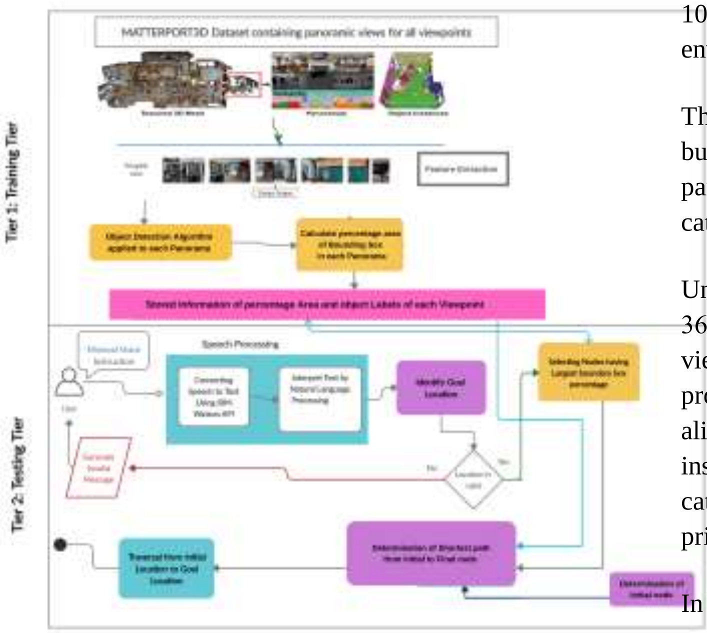
MATTERPORT3D Dataset containing panoramic views for all viewpoints
Tier 1: Training Tier
Feature Extraction
Object Detection Algorithm applied to each Panorama
Calculate percentage area of bounding box in each Panorama
Stored information of percentage Area and Object Labels of each Viewpoint
Tier 2: Testing Tier
Minimal Voice Instruction
User
Speech Processing
Converting Speech to Text Using IBM Watson API
Interpret Text by Natural Language Processing
Identify Goal Location
Location is valid
No
Yes
Generate Invalid Message
Selecting Nodes having Largest boundary box percentage
Determination of Shortest path from Initial to Final node
Determination of Initial node
Traversal from Initial Location to Goal Location

Fig 1. System Architecture

Fig 1. is the system architecture of our proposed system named Vision and Language Navigation using minimal voice instruction; it includes abstract relations with proposed features.

The architecture is a two-tier architecture consisting of training and testing tiers.

- Dataset consisting of panoramic views of all viewpoints in a given indoor environment is stored in the Matterport3D dataset.
- Object detection algorithms are applied to detect all objects visible, boundary box percentage parameter aided by a filter is used to store object information.
- Speech processing of minimal voice instruction includes use of Google Text to Speech API and NLP models. This results in identification of goal location.
- Node selection and fetching graphical information leads to determination of probable paths, eventually leading to traversal of the agent by shortest path possible.

IV. Dataset

The proposed system uses the Matterport3D dataset and investigates new research opportunities it provides for learning about indoor home environments. The dataset comprises a set of 194,400 RGB-D images captured in 10,800 panoramas with a Matterport camera in home environments.

The Matterport3D dataset provides visual data covering 90 buildings, including HDR color images, depth images, panoramic skyboxes, textured meshes, region layouts and categories, and object semantic segmentations.

Unlike previous datasets, it includes both depth and color, 360◦ panoramas for each viewpoint, samples human-height viewpoints uniformly throughout the entire environment, provides camera poses that are globally consistent and aligned with a textured surface reconstruction, includes instance-level semantic segmentations into region and object categories, and provides data collected from living spaces in private homes.

In comparison to previous datasets, Matterport3D has unique properties like RGB-D Panoramas, Precise Global Alignment & Multiple, Diverse Views of Each Surface.

V. Implementation

For simplicity, the system is represented as 3 separate modules that have been integrated together – Landmark detection Module, Natural language processing Module and Shortest path determination Module. All three modules interact with the Matterport Simulator which is hosted on a VM instance on Google Cloud.

A. Training Phase

During the training phase, we employed the landmark detection module on each node of the navigation graph of the Matterport dataset scan. At each node, there is a panorama of the visible scene which is observed and the existing objects in the scene are detected. For each detected object, we also calculate the percentage bounding box area. This helps in computation of the closest node to the goal location.

This percentage bounding box area, along with the detected label is stored at the corresponding node for later use.

B. Testing Phase

For the testing phase, we show the user an overview of the selected environment so he/she can choose where to navigate.

Once decided, the user can give a voice instruction input to the system, stating the goal location. This voice instruction is converted to text and the goal location is predicted. Common stop words like articles and conjunctions, if mentioned, are dealt with.

The predicted goal location is displayed to the user. In case of an incorrect prediction, the user can retry giving the voice instruction. In case, the specified goal location is unidentified, the user is asked to retry. Once the user is satisfied with the predicted goal location, he/she can choose to move on to the next step.

In the next step, the user is shown the original navigation graph and the initial and goal location are highlighted. Also, the nodes along the shortest path, calculated using the Djikstra's algorithm, are displayed in a different colour.

For ease of understanding, the user can choose to listen to a system-generated voice instruction for navigating from the initial location to the goal location. This instruction guides the user node-by-node to reach the destination.

Thus, using a minimal voice instruction, as short as the goal location, the user can obtain the shortest path for reaching the goal location from the initial starting node.

VI. Conclusion

The existing technologies require detailed and complex instructions to traverse to the goal location which is not very practical. The proposed system will focus on reducing the complexity of the required instructions. In the training phase, we employ a landmark detection module on the panoramas at each viewpoint in the graph generated by the Matterport Simulator. The labels of detected objects along with their bounding box area percentages are stored at each node. These are used to find the closest node to the object.

During the actual implementation, the system uses Named Entity Recognition to extract goal locations from the minimal input voice instructions. The predicted goal locations are matched with the detected label objects from the training phase. Using the Djikstra's algorithm, we find the shortest path from the initial node to the final node in the navigation graph. This approach will improve upon the graph created by the Matterport 3D dataset and drastically reduce the complexity of the required instructions to navigate to the goal location.

References

[1] P. Anderson, A. Shrivastava, D. Parikh, Dhruv Batra, and S. Lee, "Chasing Ghosts: Instruction Following as Bayesian State Tracking", NeurIPS 2019.

[2] Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. "Vision-and-language navigation

[3] Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674-3683. 2018.

[4] Liu, Peter X., Adrian DC Chan, R. Chen, K. Wang, and Y. Zhu. "Voice based robot control." In 2005 IEEE International Conference on Information Acquisition, pp. 5-pp. IEEE, 2005.

[5] Fried, Daniel, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. "Speaker-follower models for vision-and-language navigation." arXiv preprint arXiv:1806.02724, 2018.

[6] Khatib, Maher, Bertrand Bouilly, Thierry Siméon, and Raja Chatila. "Indoor navigation with uncertainty using sensor-based motions." In Proceedings of International Conference on Robotics and Automation, vol. 4, pp. 3379-3384. IEEE, 1997.

[7] Li, You, Yuan Zhuang, Haiyu Lan, Qifan Zhou, Xiaoji Niu, and Naser El-Sheimy. "A hybrid WiFi/magnetic matching/PDR approach for indoor navigation with smartphone sensors." IEEE Communications Letters 20, no. 1 (2015): 169-172.

[8] Kim, Jongbae, and Heesung Jun. "Vision-based location positioning using augmented reality for indoor navigation." IEEE Transactions on Consumer Electronics 54, no. 3 (2008): 954-962.

[9] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.

[10] Chang, Angel, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. "Matterport3d: Learning from rgb-d data in indoor environments." arXiv preprint arXiv:1709.06158 (2017).

[11] Shu-Xi, Wang. "The improved dijkstra's shortest path algorithm and its application." Procedia Engineering 29 (2012): 1186-1190.

[12] Ansh Shah, Parth Kansara, Parth Meswani, Mitchell D'silva." A review of Integrating Machine Vision and NLP for Indoor Navigation." IJARESM, Feb 2021