International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211
Volume 9, Issue 2, February -2021, Impact Factor: 7.429, Available online at: www.ijaresm.com

A Review of Integrating Machine Vision and NLP for Indoor Navigation

Ansh Shah[1], Parth Kansara[2], Parth Meswani[3], Mitchell D'Silva[4]

[1,2,3]Student, Information Technology, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India
[4]Professor, Information Technology, Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India

---

ABSTRACT

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, others have tried to integrate computer vision models into navigation algorithms. Recent advancements have integrated natural language instructions, along with computer vision techniques for navigating within the indoor environment. This paper aims to draw a comparison between the existing techniques for indoor navigation.

Keywords: Indoor Navigation, Matter port 3D, Natural Language Instructions

INTRODUCTION

A child will precisely go to the kitchen and get a glass of water when asked to do so. For robots, however, this is a complex task involving interpretation of a human instruction and using it to navigate accordingly. This complexity of intermingling speech processing with computer vision is what has been a barrier in this domain**.**

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 Matter port 3D Simulator [2] provides a large-scale reinforcement learning environment that is based on real images. 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. Alternatively, Chasing Ghosts [] applies Bayesian State tracking to propose a system, which is less dependent on navigation constraints, and is thus closer to the real-world. Each technique has its own advantages and limitations over others.

LITERATURE SURVEY

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.

Vision and Language Navigation

The system works with the Matter port 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 Matter port 3D dataset. It allowed moving between different viewpoints and adopting 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 time step t, the simulator generates an RGB image, which corresponds to what the agent observes at that particular time step.

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 time step 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 Matter port 3D simulator. Movement at each time step is based on the image observations of the agent, resulting from the movement at the previous time step. Based on this task, 21,567 navigation instructions were collected from volunteers who makes up the R2R Dataset.

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 mapped, 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 mapped, 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 Ling UNet 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.

Voice Recognition

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 modeling. 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.

Comparison and Discussions

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 discredited 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 require 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.

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.
Advantages	Separate acoustic modeling and language modeling increases the accuracy.	Success in unseen environments
Accuracy	95%	Final success rate of 53.5%.

CONCLUSION

The current approach requires 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. The proposed system will use HMMs to extract goal locations and can be framed as Bayesian State Tracking for the model that generates a

semantic spatial map of the environment, and an explicit probability distribution over alternate routes on the map. This methodology will improve upon the graph created by the Matter port 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: Interpreting visually-grounded navigation instructions in real environments." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674-3683. 2018.
[3]. 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.
[4]. 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.
[5]. 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.
[6]. 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.
[7]. 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.
[8]. 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.
[9]. 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).
[10]. Shu-Xi, Wang. "The improved dijkstra's shortest path algorithm and its application." Procedia Engineering 29 (2012): 1186-1190.
[11]. Huang, Li-Chi, Kuldeep Kulkarni, Anik Jha, Suhas Lohit, Suren Jayasuriya, and Pavan Turaga. "CS-VQA: visual question answering with compressively sensed images." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1283-1287. IEEE, 2018.