Image-Goal Navigation via Keypoint-Based Reinforcement Learning

Yunho Choi and Songhwai Oh

***Abstract*— In this paper, we tackle the problem of image-goal navigation which is a crucial robot navigation task but also a hard problem especially when there exist obstacles and limitations on field-of-view (FoV) of the camera. Conventional visual servoing approaches require depth information and camera parameters, and are susceptible to FoV loss, while previous learning-based approaches depend on the unrealistic dense reward function to train the agent with reinforcement learning. To this end, we propose a novel reinforcement learning-based approach which simultaneously utilizes self-supervised local features and global features from an observed image and a target image. The proposed method, KeypointRL, exploits keypoint matching information and generates a self-supervised reward signal which allows the agent to be easily transferred to unseen environments. The proposed model is trained on the subset of image-goal dataset in the photo-realistic Gibson dataset together with Habitat simulator, and shown to outperform baseline algorithms and generalize better.**

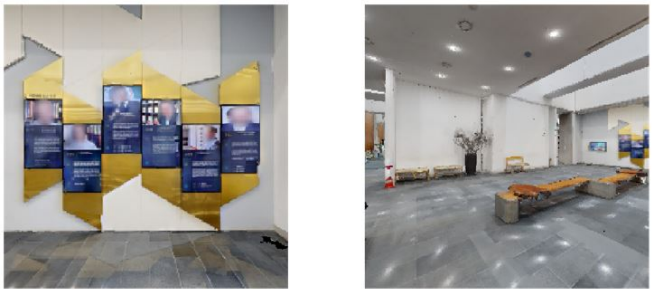
I. INTRODUCTION

Robot navigation is one of the most actively studied problem in robotics. Traditional approaches to robot navigation mainly focused on planning to reach a destination while avoiding collisions, based on a map from simultaneous localization and mapping algorithms (SLAM). However, humans do not use a map to navigate in unfamiliar environments. Humans can also carry out diverse navigation tasks such as finding an object even without a map. Recently, deep learning-based approaches are widely studied to overcome the limitations of traditional robot navigation studies and enable an human-like robot navigation. Especially, with newly-developed photo-realistic navigation simulators for embodied AI research [1], [2], many studies have shown that the learning-based approaches outperform traditional approaches such as SLAM [1], [3] and visual servoing [4]–[6].

In this paper, we focus on solving the problem of visual servoing [7] which we refer to as image-goal navigation more generally. We assume the problem setting in Figure 1. Given a RGB image at a target viewpoint (e.g. an image of a desired object), we should navigate our agents to the target viewpoint from an intial viewpoint while observing only RGB images. This problem can be relatively harder when there are many obstacles and when the agent doesn't have any overlapping field of view (FoV) between the initial viewpoint and the target viewpoint.

Y. Choi and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea. (e-mail: yunho.choi@rllab.snu.ac.kr, songhwai@snu.ac.kr).

*This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01309, Development of AI Technology for Guidance of a Mobile Robot to its Goal with Uncertain Maps in Indoor/Outdoor Environments)

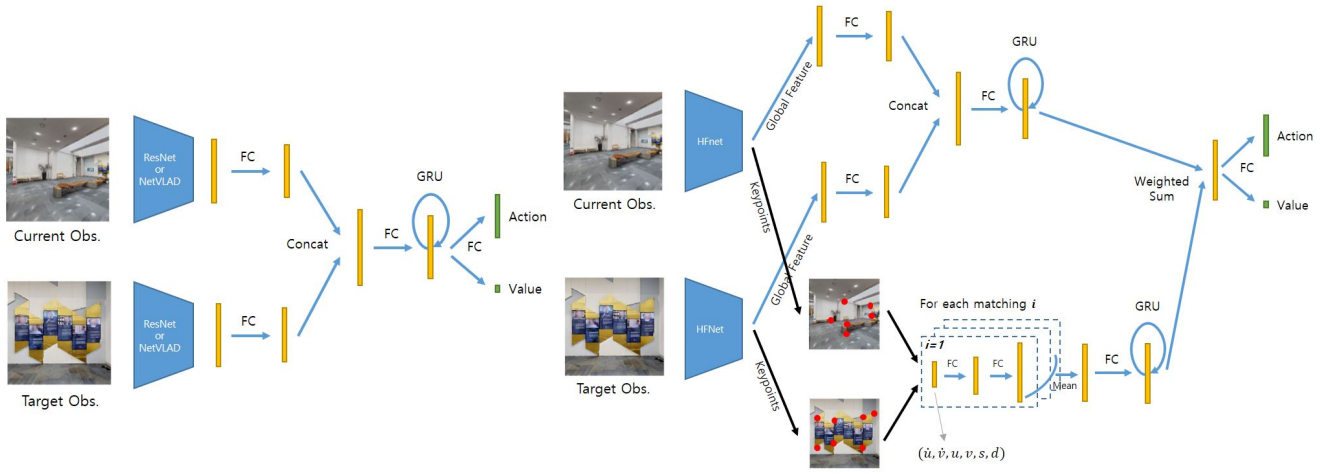


(a) Target Observation (b) Initial Observation

Fig. 1. Image-Goal Navigation: A target and an initial observations are given. Starting from the initial viewpoint, our agent's goal is to reach the viewpoint of the target observation.

Conventional approaches to visual servoing can be divided into two families, which are position-based visual servoing (PBVS) and image-based visual servoing (IVBS) [7], [8]. Both approaches make use of correspondences of local features, while PBVS approaches directly calculates the relative pose by solving the Persepective-n-Point (PnP) problem [9] and IBVS approaches calculates the robot motion using the image Jacobian [7], [8], [10]. However, both approaches require depth information of the local feature points and known camera parameters and the performance sensitively depends on the accuracy of them. They also have a constraint that the robot must have enough overlapping of FoV with the target viewpoint during navigation. To mitigate those limitations of the traditional approaches, learning-based approaches exploits expert demonstrations for imitation learning or a reinforcement learning agent [4]–[6], [10]–[13]. They commonly perform well in the simulation environment. However, they require additional information such as dense correspondences [5], [10] and additional sensor information [12] and a dense reward function based on distance to the goal, which makes their algorithms hardly transferable to unseen environments and robots in real world.

To this end, we propose a novel reinforcement learning-based method which utilizes fused feature of self-supervised local feature and global feature. Specifically, our method first extracts keypoint features from a currently observed image and target image, and summarize the result into a feature vector using a keypoint matching result between the two images. On the demand for global image features to avoid collision, we also suggest a way of fusing the global feature and the summarized local feature to make the reinforcement learning agent focus on the situationally appropriate feature. The proposed method can generate a self-supervised reward



(a) Network architecture of the image-goal navigation baselines

(b) Network architecture of the KeypointRL

Fig. 2. Network architecture of image-goal navigation baseline methods and the proposed method KeypointRL. Fully connected layer, concatenation, and gated recurrent unit are denoted by FC, Concat, and GRU respectively in the figures.

signal unlike the unrealistic dense reward signal from the previous reinforcement learning approaches, which make the learned policy easily transferred to unseen environments. In experiments, the proposed model is trained on the subset of image-goal dataset in the photo-realistic Gibson dataset together with Habitat simulator and outperform baseline algorithms.

II. Keypoint-Based Reinforcement Learning for Image-Goal Navigation

We propose a keypoint-based reinforcement learning algorithm for image-goal navigation problem, which is named as KeypointRL. Baseline reinforcement learning-based image-goal navigation algorithms, which are represented by [14], mostly extracts global image features with Siamese structure from a currently observed image and a target image as depicted in Figure 2. In addition to the global features, we propose distilling local keypoint features into the network and exploiting keypoint matching for replacing the conventional distance-based dense reward function.

A. Leveraging Local Keypoint Feature for Image Goal Navigation

In order for the agent to better understand and match the viewpoint, we propose to exploit local keypoint features of the currently observed RGB image and the target RGB image which are extracted by off-the-shelf keypoint extractor such as [15]–[18]. We first match the keypoints between the current image and the target image with the nearest neighbor algorithm. For each matching, we extract a eight-dimensional feature vector $(\dot{u}, \dot{v}, u, v, s, d)$, where $u, v, \dot{u}, \dot{v}$ are the coordinate of the keypoint in the current observation and the displacement to the matched keypoint in the target observation. We denote the keypoint descriptor compressed to a three dimensional vector and the keypoint detection score by d and s, respectively. If there isn't any matching between the two images, we mask the feature vector as a zero vector. We feedforward the feature vector into two fully connected layers and average the ouput vector across the all matchings. Then we calculate the final local feature vector with a gated recurrent unit (GRU) after a fully connected layer. Overall processing structure of the KeypointRL network is depicted in Figure 2.

B. Situational Fusion of the Global and Local Image Features

The summarized local feature vector should be fused with the global feature vector. In order to weight situationally more appropriate feature, we suggest a weighting scheme for the weighted sum of the global and local feature vector. We assume that the more different image observed and the more distant from the target viewpoint, the more we should rely on the global image feature. It is a reasonable assumption since there will be less matched keypoints from distant viewpoints and the robot should first navigate to the vicinity of the target viewpoint while avoiding collisions. From the assumption, the suggested weight w_{global} and w_{local} for the convex combination is as follows.

$$w_{global} = \Big(d_{cosine}(f_{obs}^{glob}, f_{targ}^{glob}) + (1 - n_{match}/n_{max})\Big)/2,$$

$$w_{local} = 1 - w_{global},$$

where d_{cosine} is the cosine distance, $f_{obs}^{glob}, f_{targ}^{glob}$ are global feature vectors of the current image and target image, and n_{match}, n_{max} are the number of matched keypoint pairs and the maximum number of the matched pairs, respectively. n_{max} is equal to the number of extracted keypoints $n_{keypoints}$ from the first stage. We reflect the assumption with the cosine distance which calculates how dissimilar the two images are and the the number of matched keypoint pairs.

C. Self-Supervised Image-Goal Reward Function

The previous learning-based image-goal algorithms exploit the distance-based reward function, which calculates how much the agent moved closer to the target viewpoint compared to the previous step, for every training step. This dense reward function are unrealistic since the state of the agent never contains the distance information, and makes the learned policy hard to be transferred to the real world where the distance information is not available unlike the simulator environment. Thus, we suggest a self-supervised image-goal reward function which lifts the constraint of training environment. We replace the distance-based reward with the increased number of matched keypoint pairs compared to the previous step. The reward function for KeypointRL is as follows.

$$r(s,a) = \omega_{match} * \Delta n_{match} + \mathbb{1}_{success} * r_{success}$$

, where Δn_{match} is the increased number of matched keypoints, $r_{success}$ is the success reward, and $\mathbb{1}_{success}$ indicates the success of reaching the target viewpoint.

III. Experiments

We evaluate the KeypointRL algorithm with two baseline algorithms that uses global features and the network architecture in Figure 2. Two global features are chosen, which are widely used ResNet-50 feature [19] from ImageNet pretraining [20], and NetVLAD feature [21] for place recognition. They are referred to as ResNet50 and NetVLAD in the rest of the paper. For the feature extractor of KeypointRL, we exploit the pretrained model of HF-net [16] which simultaneously outputs the self-supervised local feature [15] and the NetVLAD feature [21] so that we can save memory and computing resources.

For implementation details, we regard a navigation episode as a success when the agent's distance and azimuth between the target viewpoint are less than $1m$ and $20°$. The agent's action space is discrete and consists of {*move_forward* $0.25m$, *rotate_left* $10°$, *rotate_right* $10°$}, and we use the action noise model from the Active Neural SLAM paper [22]. For every episode, we initialize the agent's position and heading with a range of $1.5\sim5.0m$ and $30\sim90°$ away from the target viewpoint. We commonly use proximal policy optimization (PPO) [23] for training the reinforcement learning agent for fair comparison. We extract $n_{keypoints}$=48 keypoints for every observation and target observation.

We train the image-goal navigation algorithms with the eight scenes randomly chosen from the Gibson image-goal indoor dataset [2] , using Habitat simulator [1]. We evaluate the three algorithms in the test set of the same dataset. For the evaluation metric, we use success rate (SR) and success weighted by path length (SPL) [3]. As shown in Figure 3, the proposed KeypointRL outperforms the baseline algorithms and 1M training frames suffices for the successful learning. Performance in the test set is given in Table I. Despite being trained for eight rooms from the entire dataset, the proposed method generalizes well for the test set and exhibits much better performance compared to the baselines while the baselines fail to retain the performance for the training set.

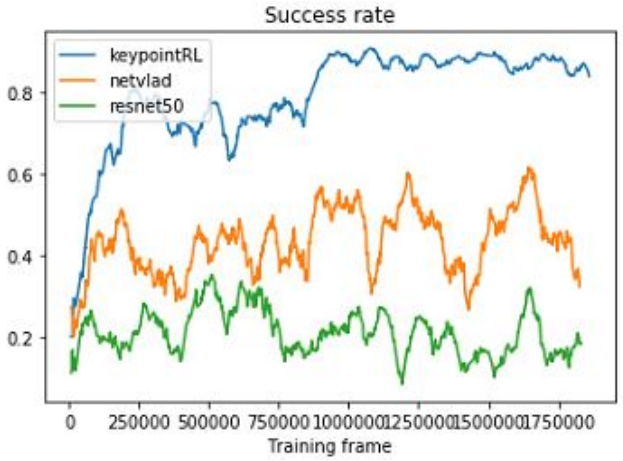


Fig. 3. Image-Goal Success rates versus the number of training frame.

TABLE I
Evaluation of Image-Goal Navigation Algorithms on a Test Set of the Gibson Image-Goal Dataset

Algorithm	Average Return	Success Rate	SPL
ResNet50	-1.7961	0.2157	0.2157
NetVLAD	1.3302	0.2941	0.2592
KeypointRL	9.5978	0.8627	0.8505

IV. CONCLUSIONS

In order to mitigate the issues of the previous visual servoing methods and the learning-based visual servoing methods, keypoint-based reinforcement learning (KeypointRL) is proposed. As validated in the experiments, KeypointRL successfully summarizes the local feature given by a keypoint detction model and fuses the summarized local feature with the global feature with the proposed weighting method. With the proposed self-supervised image-goal reward function, the navigation policy trained with KeypointRL can be readily transferred to unseen environments. We consider that learning the fusion weight for the global and local feature and deploying the algorithm on real robot are interesting directions of future work.

References

[1] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9339–9347.

[2] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv:1807.06757*, 2018.

[4] F. Sadeghi, A. Toshev, E. Jang, and S. Levine, "Sim2real view invariant visual servoing by recurrent control," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] A. X. Lee, S. Levine, and P. Abbeel, "Learning visual servoing with deep features and fitted q-iteration," *Proc. of the International Conference on Learning Representations (ICLR)*, 2017.

[6] Y. Li and J. Košecka, "Learning view and target invariant visual servoing for navigation," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[7] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[8] M. Sheckells, G. Garimella, and M. Kobilarov, "Optimal visual servoing for differentially flat underactuated systems," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[9] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[10] Y. Harish, H. Pandya, A. Gaud, S. Terupally, S. Shankar, and K. M. Krishna, "Dfvs: Deep flow guided scene agnostic image based visual servoing," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[11] F. Sadeghi, "Divis: Domain invariant visual servoing for collision-free goal reaching," *Proc. of the Robotics: Science and Systems (RSS)*, 2019.

[12] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski, and K. Alahari, "Memory-augmented reinforcement learning for image-goal navigation," *arXiv:2101.05181*, 2021.

[14] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[15] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[16] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *Proc. of Neural Information Processing Systems (NeurIPS)*, 2019.

[18] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[21] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.