Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **4**, December **2019**

Advanced SeqSLAM using Discriminative Information for Light-Rail Localization at High Frame Rate

Meng Yao, Kebin Jia*

Faculty of Information Technology
Beijing University of Technology
Beijing Laboratory of Advanced Information Networks
Beijing Advanced Innovation Center for Future Internet Technology
No.100, Pingleyuan, Chaoyang District, Beijing, China
*kebinj@bjut.edu.cn

Wanchi Siu

Department of Electronic and Information Engineering
Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

Received March 2018; revised June 2018

Abstract. *In recent years, vision-based localization technology in driving assistance system has drawn much attention. In this paper, an Advanced SeqSLAM method is proposed to solve the problem of localization due to the high similarity of scenes in high-accuracy scene matching of light rail system. In this method, salient regions with discriminative information are extracted from high-similarity frames of reference sequences by off-line processing, and binary feature descriptors are generated in these regions to improve the speed and precision of scene matching. Compared with the local features, the error of the proposed scene matching method is reduced by 31.43% and the computation time is reduced by 94.22% in the Hong Kong MTR dataset. Compared with the scene tracking algorithm of SeqSLAM, the precision of scene tracking based on proposed binary features in salient regions is increased by 9.84% compared without significant increase of running time in the Nordland dataset. The experimental results show that the proposed method improves the performance of the light rail localization.*
Keywords: Vision-based localization; Scene tracking; Salient region detection; Binary feature extraction

1. **Introduction.** In recent years, advanced driver assistance systems (ADAS) are widely used in vehicle scheduling systems to improve the safety. As an important part of ADAS, the localization module is the basis of other function modules. Due to the complex environments including tunnels, urban canyons formed by skyscrapers and even inside of the buildings, the instable signal in the ADAS system for light rail based on Global Position System (GPS) poses a huge safety risk for train driving and scheduling. Therefore, the localization technology based on visual information has become an active research area [1, 2].

Vision-based localization system collects visual information during vehicle travelling and transforms it into topological map [3], which is stored in the database. The nodes and edges contained in the topological map represent the defined scenes and the relationships between scenes, respectively. When the vehicle enters the same scene again, the

localization system locates the current position based on the current frame taken by the camera using scene matching to find the most similar node/scene in the topology map. In the light rail localization system, the topology map can be simplified to one-dimensional scene chain, as shown in Fig.1. Meanwhile, the location information can be obtained by the route-based scene tracking algorithm [4].

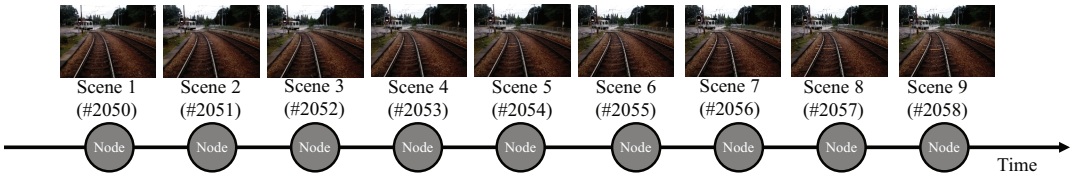


FIGURE 1. Topologic map for light rail localization system

The scene matching based on visual information may suffer from condition changes such as moving objects and illumination changes. Therefore, laser radar [5], infrared sensor [6] and stereo camera [7] are widely used in localization system to obtain the stable information of scene. However, these methods rely on special sensors and can hardly be migrated to mobile platforms. Therefore, monocular camera-based visual localization is still a hot area of research [8, 9].

Conventional visual feature descriptors, such as Scale-Invariant Feature Transform (SIFT) [10] and Speeded-Up Robust Feature (SURF) [11] are used to generate the descriptors of scenes [12, 13, 14]. The machine learning-based methods extract the stable features of the scene [15, 16, 17], or remove illumination-sensitive components [8, 9, 12, 13]. In recent years, with the success of deep learning methods, Convolutional Neural Networks (CNNs) [18] have been applied to the vision-based localization system to obtain stable scene information [14, 19]. Scene change learning [17, 20, 21, 22] predicts the condition changes to match scenes under different conditions. However, these learning-based methods require a large number of video sequences and manual calibration to generate training data. Therefore, the localization system based on single reference sequence still faces many challenges [7].

The accuracy of vision-based localization is determined by the frame rate of reference sequence. The reference sequence at higher frame rate records more location information. Therefore, location information with high accuracy can be obtained through frame-by-frame matching. However, due to the high similarity of the scenes in high-frame-rate railway sequence, such as 25 frame per second (fps) shown in Fig.1, the reference sequence is down sampled to 1 fps in the time domain [4, 12] to increase the match rate but reduce the accuracy. Therefore, vision-based localization with high frame rate reference sequence is still an important bottleneck.

In order to solve the problems of large visual data training, low localization accuracy and high computational complexity, this paper proposed an Advanced SeqSLAM method with salient region detection and binary feature extraction for high-accuracy scene matching based on single monocular sequences. Compared with other methods, the proposed method has the following innovations: 1) a salient region detection method is proposed for single monocular reference sequence to distinguish continuous similar frames (as shown in Fig.1); 2) a salient region-based binary feature is proposed to accelerate visual feature extraction for scene matching, which meets the requirement of real-time scene tracking system; 3) an Advanced SeqSLAM scene tracking approach based on discriminative information is designed for high accurate light rail localization.

2. **Method.** This part contains the two parts of our system: discriminative information learning in offline module and discriminative information based scene tracking in online module. As shown in Fig.2, the salient regions in each reference frame are labeled firstly. The binary patterns of these salient regions are generated, which are used to extract the binary features of the reference frames and current frames.

In the online module, for each current frame, a series of candidate matching reference frames are retrieved by the SeqSLAM method in the offline module. The best matched reference frame for the current frame is identified by the binary feature verification to obtain the current location of the train.

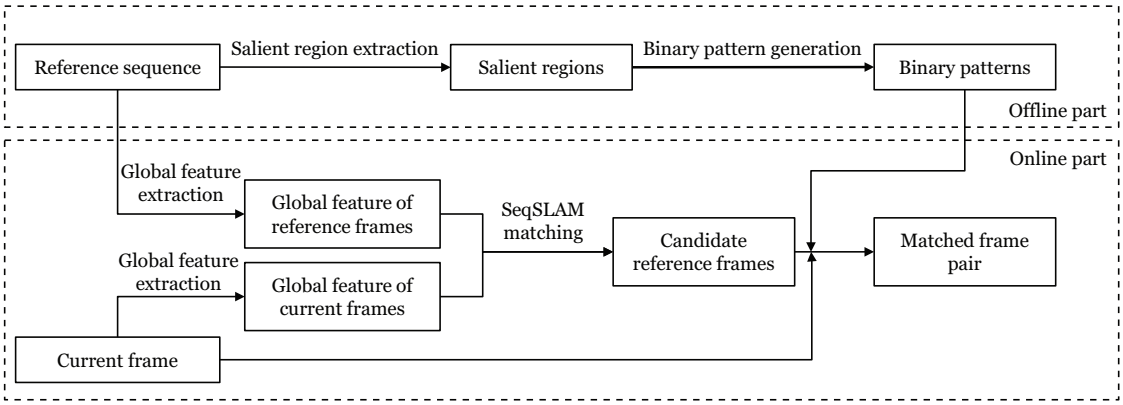


FIGURE 2. The framework of the proposed method

2.1. **Discriminative information learning.** The expected salient regions contain the difference between high-similarity frames. To reduce the computational complexity and improve the stability of the algorithm, we first establish the region of interest (ROI) in the frame for further key region detection.

For vision-based localization system, three categories of useless areas should be removed from the ROI in light-rail sequence frame. The removed areas include a rectangular area with temporary occlusion caused by other trains, the track area without significant change over time and the blur area near the boundary of the frame.

The salient regions in the video frame consist of pixels with higher discrimination power, and the saliency scores are used to measure this power of the pixels. The saliency score of a pixel reflects the difference between that pixel and the corresponding pixels in same location of another frames. The pixels in a video frame ft at a certain moment t are denoted as $p(x, y, f_t)$. The set of neighbor frames is denoted as $\mathbf{F}(t)$. The visual information of $p(x, y, f_t)$ is represented by the Histogram of Oriented Gradients (HOG) feature [23] in the surrounding area. The saliency score S of $p(x, y, f_t)$ can be calculated by equation (1).

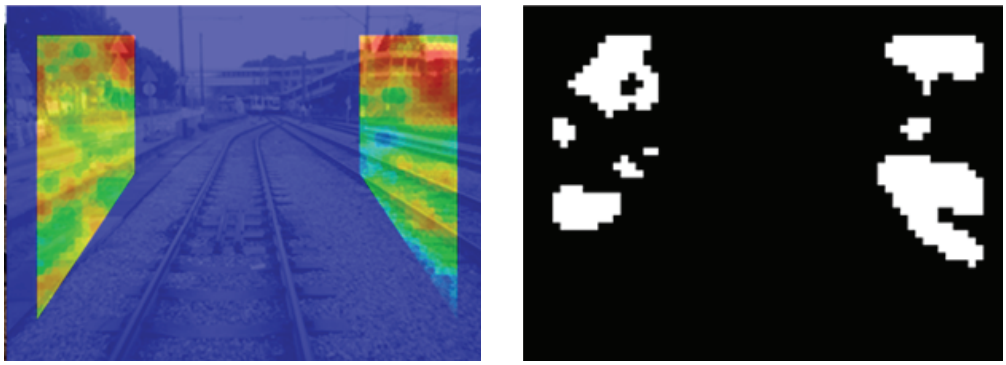
$$S(x, y, f_t) = \frac{1}{T} \sum_{f_i \in \mathbf{F}(t)} \|\vec{D}(x, y, f_t) - \vec{D}(x, y, f_i)\|_2 \tag{1}$$

where the $\vec{D}(\cdot)$ is the HOG feature vector of pixel $p(\cdot)$. The $\mathbf{F}(t)$ is defined by equation (2).

$$\mathbf{F}(t) = \left\{ f_i | i \in \left[t - \frac{T}{2}, t + \frac{T}{2} \right], i \neq t \right\} \tag{2}$$

where T is the number of the neighbor frames in $\mathbf{F}(t)$.

After calculating the saliency scores of all the pixels in the ROI, the salient regions are obtained by grouping the pixels which have higher scores than the predefined threshold. Fig.3(a) shows the saliency scores of the pixels within ROI of a frame and the salient regions extracted from the frame. The red area represents a region of high significance, and the blue color in turn. When the value of threshold is 1.05 times the average value of the saliency scores in the ROI, Fig.3(b) contains two salient regions while the small regions with the bounding box smaller than 40×40 are easily disturbed by noise and not regarded as the salient regions.



(a) (b)

FIGURE 3. Example of saliency score and salient regions. (a) Heat-map of saliency score in ROI. The pixels with higher saliency score are drawn in red, while blue areas have lower saliency score. (b) The extracted salient regions.

Compared with the conventional feature description, the binary features represented by BRIEF [24] and ORB [25] have less computational complexity in the extraction and matching stage, and are widely used in real-time systems. However, these binary feature descriptors are designed for small regular regions, which means not suitable for large and irregularly shaped regions. In this section, we generate specific binary patterns for salient regions in the reference sequence to calculate the binary features of the reference frame and the current frame when they are in the matching procedure.

The binary feature vector consists of cascaded bits, each of which reflects the intensity relationship of a certain pixel pair in the feature description area, as shown in equation (3).

$$\tau(p_1, p_2) = \begin{cases} 1, & I(p_1) > I(p_2), \\ 0, & otherwise, \end{cases} \tag{3}$$

where $I(\cdot)$ is the pixel intensity. $t(\cdot)$ is the binary test.

For one reference frames f_t, create its pixel set $P(f_t)$, which contains all the pixels in the salient regions. The binary pattern $\mathbf{H}(f_t)$ is established by randomly selecting N pixel pairs in $\mathbf{P}(f_t)$. The binary descriptor $B(f_t)$ can be generated by equation (4).

$$B(f_t) = \sum_{0 \le i \le N} 2^i \tau(p_m, p_n) \tag{4}$$

where p_m , $p_n \in H(f_t)$. N is the number of the pixel pairs in $H(f_t)$.

The pixel pairs in $H(f_t)$ can be divided into two categories. The intra-pair contains two pixels coming from the same salient region which records the local visual information of the region. Another type, called inter-pair, contains two pixels coming from different regions. Therefore, the pixel pairs record the relative position between these salient regions.

2.2. **Advanced SeqSLAM scene tracking.** For each current frame, the tracking module in the online phase retrieves the best matching frame in the reference sequence using the SeqSLAM. However, due to the high frame rate of the reference sequence, the SeqSLAM fails to identify the most accurate matching frame and return a series candidate frames with similar appearances. The proposed advanced SeqSLAM verifies these candidates with learning-based binary feature to obtain the best matching reference frame.

Denote the current frame as $f_{C,t}$. $\mathbf{Q}_t$ is the candidate frame set returned by SeqSLAM. The extracted binary feature of frame $f_{R,i}$ in $\mathbf{Q}_t$ is denoted as $B(f_{R,i})$. The Hamming distance is used to measure the similarity between $f_{C,t}$ and $f_{R,i}$, which is denoted as $H(B(f_{C,t}, B(f_{R,i}))$. The best matching frame $fmatched$ for $f_{C,t}$ can be identified with equation (5).

$$f_{matched}(f_{C,t}) = \arg\min_{f_{R,t} \in \mathbf{Q}_t} \left(H(B_{f_{C,t}}, B_{f_{R,i}}) \right) \tag{5}$$

where $H(\cdot)$ is the Hamming distance function.

3. **Experimental results and discussions.** In order to evaluate the validity and performance of the proposed salient region and the binary feature extraction method, we collect the results of two experiments using the Hong Kong MTR dataset and a publicly available Nordland dataset and give some discussions in this section.

3.1. **Dataset.** The two video sequences in the Hong Kong MTR data set were captured by our smartphone installed in light rail vehicles with a video resolution of 640×480 and a frame rate of 25 frames per second. Due to the different collection time, the illumination condition and the train speed are all different in these two sequence. All frames are manually calibrated. The Nordland database contains four sequences collected in four seasons with a video resolution of 1920×1080 and a frame rate of 25 frames per second. In this paper, 6000 frames in summer and fall are used as training and testing data and downsampled to 640×480. The two sequences keep running at the same speed. Therefore, the frames with same index number were collected from the same location.

3.2. **Evaluation for salient region.** In order to evaluate the validity of the salient region proposed in this paper, we compared the accuracy of HOG-based scene matching with salient regions and other 3 methods without salient regions: 1) a global HOG feature that uses a HOG feature to describe the entire video scene; 2) the local HOG feature that divides the frame into 40×40 macroblocks and calculates HOG feature vectors in each of the macroblocks respectively; 3) the HOG feature vector of macroblocks located in the region of interest is calculated. 4) HOG feature vectors are calculated in the salient regions proposed in this paper.

We matched the single frame scene in the reference sequence with the continuous scene with high similarity in the current sequence within Hong Kong MTR dataset, and used the offset between the matching result and the artificial calibration result as the matching deviation. The average calculation time and matching offset of the four methods are shown in Table 1.

TABLE 1. Computational time and matching offset of matching.

Method	Matching offset (frame)	Computational time (s)
Global HOG feature	15.24	0.0593
Local HOG feature	2.10	62.4205
HOG feature in ROI	2.26	13.5960
HOG feature in salient region (proposed)	1.44	3.6058

The scene matching based on global HOG features is the fastest, but the matching offset with value of 15.24 makes it can be hardly used in the real application. Although the matching offset of scene matching based on the local HOG features drops to 2.10 frames, the huge computation time can not meet the requirement of the practical system. On the other hand, the computation time of scene matching based on the ROI decreased by 78.22%, but the matching offset was increased by 0.16 frame. Compared with the local HOG feature method, the matching offset is reduced by 31.43% and the matching time is reduced by 94.22% based on the scene matching method proposed in this paper.

The experimental results show that the global HOG features only extract the rough visual feature of the whole video frame. Although the computational complexity is low, the matching offset is too large due to the lack of the detail information of scene. The local HOG features based method can record both the details and global information of the scene. However, the high complexity of HOG feature calculation and feature matching makes this method unable to be used in the practical system. The ROI based method reduces the number of macroblocks used to calculate HOG features, thus greatly reducing the time complexity of scene matching. With the salient region based method, the scene matching module only calculates the region with the most discriminative information in the scene. Meanwhile, it reduces the computational complexity of scene feature calculation and scene matching and reduces the noise interference caused by non-critical regions to scene matching.

3.3. **Evaluation for Advanced SeqSLAM scene tracking.** The proposed Advanced SeqSLAM scene tracking method was tested in Nordland dataset. The reference sequence was the fall sequence, and the summer sequence was used as the current sequence. As a novel route-based scene matching algorithm, SeqSLAM [4] is widely used in path-based scene tracking algorithms [26, 27] and compared with proposed method. We also compared our approaches with a state-of-art LDB feature [28]. The matching offset between the actual result of the scene matching and the ground truth of less than 3 frames is considered as the correct match and vice versa as the wrong match. After counting all the correct matches and wrong matches, the matching precision can be calculated using equation (6). The average of matching offset between the actual and ideal results was also calculated to measure the performance.

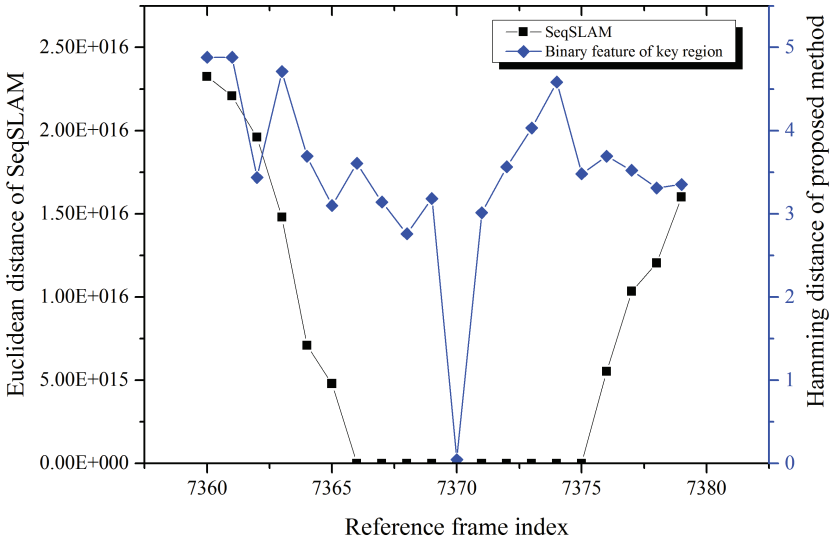
$$Precision = \frac{correct\ match}{correct\ match + wrong\ match} \times 100\% \tag{6}$$

Table 2 shows the comparison of precision and computational time between proposed method and other methods. Compared with SeqSLAM, the scene tracking precision of proposed method is improved by 9.84% and the matching offset is reduced by 39.79% without significantly increasing the time cost. The proposed method also shows the

better performance than the LDB-based method with increasing the precision by 2.68% and reducing the offset by 38.66%.

TABLE 2. Precision and computation time of scene tracking.

Method	Precision	Matching offset (frame)	Time (ms)
SeqSLAM	89.56%	1.3652	53.23
LDB	96.72%	1.3400	51.55
Proposed method	99.40%	0.8220	54.82



(a) Distance distributions of SeqSLAM and proposed method Current frame C#7370 ?

(b) Reference frame R#7366–R#7375

FIGURE 4. The distribution of matching distance of SeqSLAM and proposed method.

Fig.4(a) shows the matching distance distribution of the current frame C#7370 in the neighborhood of ground truth (R#7370 reference frame). In this figure, the horizontal axis is the reference frame number, the left vertical axis is the matching distance of the SeqSLAM method, and the right vertical axis is the Hamming distance based on the proposed Advanced SeqSLAM scene tracking method in this paper.

In the SeqSLAM matching result, the matching distances between current frame C#7370 and 10 reference frames, including R#7366 and R#7375, are 0, as shown by the black line in Fig.4(a). In contrast, the binary feature proposed in Advanced SeqSLAM reaches the minimum matching distance at the ideal matching result (horizontal axis 7370), as indicated by the blue line.

Fig.4(b) shows the 10 reference frames with the matching distance of 0 in SeqSLAM. These 10 reference frames contain high similar scenes. This shows that the global features used by SeqSLAM reserve only the rough information about the scene and can hardly distinguish between these scenes with high similar appearance. The proposed binary features are generated in the salient regions, while the salient regions contain the most prominent visual information of the frame relative to the neighbor frames. The experimental results show that the binary feature proposed in Advanced SeqSLAM preserves these salient features and have better performance in scene tracking at high frame rate.

4. **Conclusions.** This paper presents an Advanced SeqSLAM for high-accuracy real-time scene matching with salient region detection and binary feature extraction. Aiming at the problem of high similarity of continuous scene at high frame rate sequence, this method analyzes and extracts the salient regions containing the discriminative information of scenes frame by frame. Meanwhile, in order to meet the requirement of real-time system, the binary features are extracted for the irregular salient regions. The experimental results in Hong Kong MTR dataset and Nordland dataset show that the proposed salient region and binary feature can improve the precision of scene matching while ensuring the efficiency of scene tracking system.

Acknowledgment. This work was supported by the National Natural Science Foundation of China [No. 61672064]; and the Beijing Natural Science Foundation [No. KZ201610005007].

REFERENCES

[1] S. Lowry, N. S nderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, Visual place recognition: A survey, *IEEE Transactions on Robotics*, vol.32, no.1, pp.1-19, 2016.

[2] F. Dornaika, A. Assoum, and A. Moujahid, Place Recognition Using Image Retrieval with Covariance Descriptors, *Electronic Imaging*, vol.2016, no.10, pp.1-6, 2016.

[3] H. Shatkay, L. P. Kaelbling, Learning geometrically-constrained hidden markov models for robot navigation: Bridging the topological-geometrical gap, *Journal of Artificial Intelligence Research*, vol.16, no.1, pp.167-207, 2002.

[4] M. J. Milford, G. F. Wyeth, SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights, *IEEE International Conference on Robotics and Automation (ICRA)*, Saint Paul, MN, USA, pp.1643-1649, 2012.

[5] C. Mcmanus, P. Furgale, and T. D. Barfoot, Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders, *Robotics and Autonomous Systems*, vol.61, no.8, pp.836-852, 2013.

[6] W. Maddern, S. Vidas, Towards robust night and day place recognition using visible and thermal imaging, *RSS 2012: Beyond laser and vision: Alternative sensing techniques for robotic perception*, Sydney, Australia, 2012.

[7] C. Linegar, W. Churchill, and P Newman, Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera, *International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, pp.787-794, 2016.

[8] P. Ross, A. English, D. Ball, B. Upcroft, G. Wyeth, and P. Corke, A novel method for analysing lighting variance, *Australian Conference on Robotics and Automation*, Sydney, Australia, 2013.

[9] P. Ross, A. English, D. Ball, and P. Corke, A method to quantify a descriptor's illumination variance, *Australian Conference on Robotics and Automation*, Melbourne, Australia, 2014.

[10] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol.60, no.2, pp.91-110, 2004.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, Speeded-up robust features (SURF), *Computer Vision and Image Understanding*, vol.110, no.3, pp.346-359, 2008.

[12] P. Corke, R. Paul, W. Churchill, and P. Newman, Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localization, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, pp.2085-2092, 2013.

[13] C. McManus, W. Churchill, W. Maddern, A. Stewart, and P. Newman, Shady dealings: Robust, long-term visual localisation using illumination invariance, *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, pp.901-906, 2014.
[14] Z. Chen, O. Lam, A. Jacobson, and M. Milford, Convolutional neural network-based place recognition, *Australasian Conference on Robotics and Automation*, Victoria, Australia, 2014.
[15] P. Neubert, N. S nderhauf, and P. Protzel, Superpixel-based appearance change prediction for long-term navigation across seasons, *Robotics and Autonomous Systems*, vol.69, pp.15-27, 2015.
[16] M. Milford, E. Vig, W. Scheirer, and D. Cox, Vision-based simultaneous localization and mapping in changing outdoor environments, *Journal of Field Robotics*, vol.31, no.5, pp.780-802, 2014.
[17] S. Lowry, M. J. Milford, Supervised and unsupervised linear learning techniques for visual place recognition in changing environments, *IEEE Transactions on Robotics*, vol.32, no.3, pp.600-613, 2016.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Proceeding NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp.1097-1105, 2012.
[19] N. S nderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, On the performance of convnet features for place recognition, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, 2015.
[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. S sstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.34, no.11, pp.2274-2282, 2012.
[21] S. Lowry, M. J. Milford, and G. F. Wyeth, Transforming morning to afternoon using linear regression techniques, *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, pp.3950-3955, 2014.
[22] S. Lowry, G. Wyeth, and M. Milford. Unsupervised online learning of condition-invariant images for place recognition, *Australasian Conference on Robotics and Automation*, Melbourne, Australia, 2014.
[23] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, USA, pp.886-893, 2005.
[24] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, Brief: Binary robust independent elementary features, *European Conference on Computer Vision*, Heraklion, Crete, Greece, pp.778-792, 2010.
[25] E. Rublee, V. Rabaud, and K. Konolige, ORB: An efficient alternative to SIFT or SURF, *2011 IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, pp.2564-2571, 2011.
[26] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, Simultaneous Localization and Mapping: A Survey of Current Trends in Autonomous Driving, *IEEE Transactions on Intelligent Vehicles*, vol.2, no.3, pp.194-220, 2017.
[27] P. Kim, B. Coltin, and O. Alexandrov, and H. J. Kim, Robust visual localization in changing lighting conditions, *IEEE International Conference on Robotics and Automation*, Singapore, pp.5447-5452, 2017.
[28] X. Yang, K. Cheng, Learning optimized local difference binaries for scalable augmented reality on mobile devices, *IEEE Transactions on Visualization and Computer Graphics*, vol.20, no.6, pp.852-865, 2014.