Vision-based Localization with Monocular Camera for Light-rail System

Kebin Jia[1,2,3,*], Tingxian Wang[1,2,3] and Meng Yao[1,2,3]
[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2] Beijing Laboratory of Advanced Information Networks, Beijing 100124, China
[3]Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China
[*]E-mail: kebinj@bjut.edu.cn Tel: +86-10-67392529

***Abstract*— As an emerging localization method, vision-based localization methods have been widely used in vehicle safety system. By considering the practical requirements like the high accuracy, real-time performance and easy installation, we design a localization system for urban light rail based on the monocular camera. This system is divided into two parts: offline and online. To solve the problem of scene matching with high similarity, we proposed a new scene recognition method based on local key regions and key frames. This method not only guarantees the precision of scene matching but also satisfies the real-time requirement of the system. The offline module uses the unsupervised method to extracts the key region with discriminative information from high-similarity frame of reference sequences and selects the key frame based on it. The online module can quickly match the current frame and reference frame within the retrieval range provided by the key frame, by calculating the binary feature with a low correlation in the key regions. While meeting the high-precision needs of the light rail system, it significantly improves real-time performance. This paper uses both the public test dataset in Nordland and the challenging Hong Kong light rail dataset. The experiment results show that the proposed method can accomplish rapid and accurate light-rail localization at high frame rate. The precision can reach more than 90% in extreme situations such as large-area scene occlusion.**

I. INTRODUCTION

Nowadays, advanced driver assistance systems (ADAS) are widely used in vehicle scheduling systems to improve the safety and efficiency. As an important part of ADAS, the localization module should meet the requirements of real-time and higher accuracy. Currently, the Global Positioning System (GPS) is widely used in the vehicle localization system, and its accuracy can reach about 5 m, which applies to large-scale localization systems, such as the common intercity train dispatching system. Different from ordinary vehicles, the operating environment of urban light rail is often complex. The unstable signal in the ADAS system for light rail based on the Global Position System (GPS) poses a huge safety risk for train driving and scheduling [1].

In recent years, visual information plays an important role in the localization system and is widely used in vehicle and mobile robot navigation system [1-3]. Vision-based localization system continuously collects visual information during vehicle travelling and transforms it into topological map [4], 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. Meanwhile, the location information can be obtained by the route-based scene tracking algorithm.

As matter of fact, scene matching is often interfered with condition changes such as illumination changes and partial occlusions. Therefore, infrared sensor [5], lidar [6], stereo camera [7] are widely used in localization system to extract the stable features of the scenes which are unaffected by drastic environmental changes. However, compared to the monocular camera, these methods rely on special sensors and do not have the advantages of low maintenance costs, and non-susceptible to external signal interference, etc. Hence, monocular camera-based visual localization system is still the hot area of research [8].

Since a train always runs in the wild where the scenes are always similar in a period, matching each live frame with reference frames in memory cannot tell the current location exactly with high confidence. Therefore, we proposed a binary feature extraction method based on key regions by off-line processing and accomplish a real-time localization system for light rail with monocular camera. The system is based on the following innovations: (i) the proposed unsupervised key region and key frame extraction method, which is suitable for reducing the computational complexity of scene matching; (ii) a learning-based method to identify the binary feature of key regions which can be used in real-time.

The paper is structured as follows. A summary of related works is given in section II. In section III, the details of the proposed method are described. The experimental results and some further discussion are given in section IV. Finally, some conclusions are drawn in section V.

II. RELATED WORKS

Feature points are widely used in traditional algorithms to locate the stable feature, such as SIFT [9], SURF [10], and FAST [11]. With these feature points, local features can be extracted with different kinds of description methods, for example, SIFT [9], SURF [10], BRIEF [12], and ORB [13]. Feature points can be obtained in different formats, including difference images of Gaussian [9], responses of Fast-Hessian detector [10], or learning-based corner detectors [11]. Many algorithms were successfully developed in many matching systems, and they are good, but mainly focus on the local information within a small patch around the feature points, which would be unstable in the scene matching and recognition systems with huge illumination changes or viewpoint changes [14].

Recently, some feature extraction methods on the problems of visual-based localization systems were proposed. Scene signatures [14] is a viable way to extract stable features of a place with a dataset covering most extreme appearance changes. The feature detector for one place proposed in this paper was trained with abundant data captured in different conditions, such as sunny, raining, snowing, and deep darkness. Han et al. [15] proposed a Shared Representative Appearance Learning (SRAL) method, which integrates multiple image features and implements a vehicle localization algorithm based on this feature. Carlevaris-Bianco et al. [16] used 3 million training samples to track the stable features in the images that did not change with time.

The neural network maps the training data into a lower-dimensional feature space which is more robust than the hand-designed features. Focusing on the viewpoint-invariance and condition-invariance, Convolutional Neural Network (CNN) [17] was used to train a landmark detection to identify stable features [18]. In this algorithm, the Edge Boxes [19] provides the object-like regions as the candidates of the landmark. The potential landmarks are extracted from these candidates with the feature generated by CNN. Arroyo et al. [20] design Convolutional Neural Network for Visual Topological Localization (CNN-VTL) method, which uses a large amount of training data to obtain scene features for vehicle localization. This kind of sample-based learning method requires to collect a large amount of scene information and conduct manual calibration, so the vehicle localization algorithm based on a single reference sequence still faces many challenges [21].

Our work is significantly different from former approaches. We just use a single reference sequence captured by a monocular camera to extract the discriminative information with an unsupervised method independently. The learning-based method just requires to extract the binary feature of key regions in the real-time matching procedure. The most prominent advantage of our system is that neither requires a large training set nor relies on any special sensors.

III. METHOD

The detail of the proposed method will be given in this section. As shown in Fig. 1, the vision-based localization system for light rail contains offline and online two parts. To solve the problem of the high similarity of reference frames, the offline module extracts the key region with discriminative information for each reference frame and the frames having a outstanding appearance in video are regarded as key frames. The key regions in each frame are labeled firstly. The binary patterns of these key regions are generated, which are used to extract the binary features of the reference frames and current frames. In the online module, when the light rail is running on the same path as the reference sequence, we use the SeqSLAM method [23] to obtain a series of candidate matching reference frames within the retrieval range provided by the key frame. The best-matched reference frame for the current frame is identified by the binary feature verification to obtain the current location of the light rail.

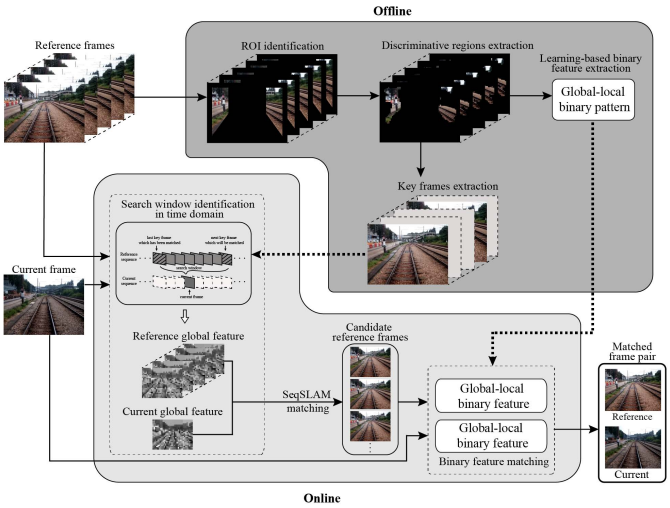


Fig. 1 The framework of the proposed light-rail localization system.

A. Key Region Detection

The expected key regions contain the difference between high-similarity frames. We first establish the region of interest (ROI) in the frame for further key region detection. The frames in the vision-based localization contain 3 types of useless regions, including moving objects, railway track, and blur region near the boundaries. A rectangle with 300 pixels in the center of the frame was removed because of the temporary occlusion of the front train. The triangle regions containing the railway which is useless for localization were discarded. A margin of 40 pixels at the boundary of the frame is not in the ROI because of serious blur and distortion.

All the pixels in the ROI record the overall information of the scene. This global information can locate the approximate position of the vehicle, such as the stop platform or the driving section. Only specific regions contain significant information that helps provide more accurate location information for the vehicle, known as the key region. The discriminative score is used to measure the saliency degree of the region within the frame. The higher the score is, the more significant the region will be. In this paper, the sliding

window is used to sample the ROI for each frame to calculate the regional discriminative score.

For instance, let us denote the query frame that we want to extract key regions as f_t. The neighboring frames around the query frame consist of a set of F frames; for example, 4 frames as shown in Fig. 2. When the sliding window moves to the position (x, y), as shown with red rectangle, patch $R(x, y, f_t)$ within this window is compared with co-located patches $R(x, y, f_{t'})$ in other frames in the reference set. The

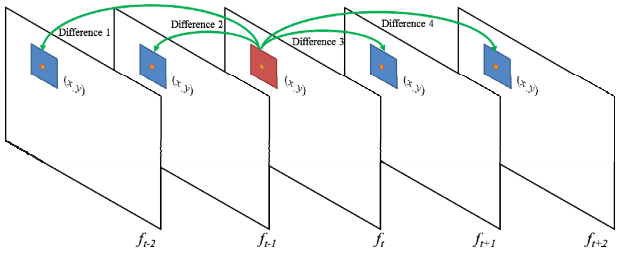


Fig. 2 Discrimination power computed by summing Euclidean distance

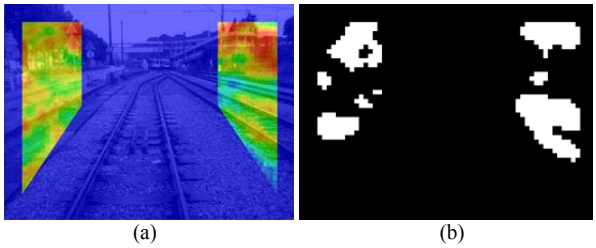
discrimination score of $R(x, y, f_t)$ can be computed by summing these differences, as shown in (1).

$$S_{R(x,y,f_t)} = \frac{1}{N} \sum_{f_{t'} \in F, t' \neq t} D(R(x,y,f_t), R(x,y,f_{t'})) \qquad (1)$$

where $D(R_A, R_B)$ is the function to compute the difference between patches R_A and R_B. $R(x, y, f_{t'})$ is the reference patch in frame $f_{t'}$. N is the number of the frames in the reference set. N=4 is used in Fig. 2 $S_{R(x, y, ft)}$ is the summed geometric distance.

The difference between two image patches can be measured by the summed absolute pixel intensity differences or the Euclidean distances of feature vectors. Our system uses a Histogram of Oriented Gradients (HOG) [22] features to calculate image patch differences to avoid the influence of illumination change. The discriminative score reveals the saliency of the image patch. As shown in Fig. 3(a), the discriminative power of each pixel was drawn with different colors. Red regions have a higher discrimination score than blue regions.

Pixels that have higher distance than a predefined threshold T_k, is regarded as within the key region. Hence key regions can be obtained by grouping those pixels together. Fig. 3(b)



(a) (b)

Fig. 3 The discrimination power of ROI and the selected key region

shows a sample of extracted key regions. The white regions are the extracted key regions.

B. *Key Frame Extraction*

The discriminative score of a frame can be calculated by summing the score of all key regions within this frame. The key frames are expected to contain specific information in a sequence of video over some time. Therefore, the key frames can be defined as the frames which have higher discrimination score compared with not only remote frames in the sequence but also neighboring frames, so that the online matching module can get a high-confidence matching.

The key frame extraction method consists of two steps. Firstly, video frames with the local maximal discriminative score are extracted and these frames are sorted in descending order according to the discriminative score. Secondly, the former N_k frames are taken as the key frame.

C. *Learning-based Binary Features Extraction*

The extracted key regions method proposed in this paper reduce the area of scene matching and computational complexity. On this basis, the system can focus on some local visual information with high resolution to identify an accurate location. The local visual information can be extracted by the statistic method, such as HOG or SIFT feature. However, the high computational complexity makes these methods not be efficient for the real-time system. At the same time, this kind of floating-point feature descriptor using Euclidean distance to calculate feature similarity is responsible for the matching process very time-consuming. To improve the efficiency of feature extraction and matching process, a variety of binary feature descriptors come up, such as BRIEF and ORB, which are designed mainly for general local feature description within a rectangle patch. Especially, the ORB features are based on corner point description, and binary feature extraction mode is obtained by learning method. In this paper, a learning-based binary descriptor is proposed, which contained higher discriminative information for irregular key regions descriptor. A novel saliency analysis and greedy algorithm are used in this approach.

The descriptors can be obtained by cascading binary comparisons result of a series of pixel pairs. A powerful binary feature means to have pixel pairs which can make the current query frame outstanding from its neighboring frames. The proposed method is used to extract the pixel pairs with the strongest discrimination power. A discrimination score is used to evaluate the power of the pixel pairs to separate the current query frame from others. The discriminative score of a pixel pair P can be calculated with (2).

$$S(P, F_q) = \sum_{i=0}^{M} \left(D(P, F_i) - D(P, F_q) \right) \qquad (2)$$

where $S(P, F_q)$ is the pixel difference of P in query frame F_q, and $D(P, F_i)$ is the pixel difference of pixel pair P in the i^{th} neighboring frame. M is the quantity of the neighboring frames near the query frame.

All pixel pairs were sorted with the discrimination scores in descending order. And we can choose the first N pixel pairs to identify the query frame. However, the comparison results of

similar pixel pairs always have high correlation although the discrimination score of them are both high. For instance, if we choose the pixel pair $P((x_1, y_1), (x_2, y_2))$ with high discrimination score, the neighboring pair $P((x_1+1, y_1+1), (x_2+1, y_2+1))$ may have the similar high score and be chosen as well. The information of binary descriptor generated by choosing the pixel pair based on the discrimination may be reduced. Therefore, the pairs with low correlation should further be identified.

Principal Component Analysis (PCA) is one of the most useful methods to identify the principal dimensions of the data. However, it is not suitable for our system, because there are only a few training samples, which cannot provide sufficient data for PCA to extract the desired number of dimensions. Therefore, the PCA method cannot be used to identify point pairs with low correlation. A greedy method is used in our approach to check all possible pixel pairs in the key region, which can identify good pixel pairs with high discrimination scores and low correlations.

In this paper, the greedy algorithm based on cross-correlation numbers can extract high-quality binary features. The training set using in the greedy algorithm consists of the current frame and several frames in its neighborhood. First of all, the pixel difference of all possible pixel pairs in both the query frame and neighboring frames are calculated to establish the training matrix ***T***. Assuming that the current frame contains K pixels and M adjacent frames in the key region, the number of all pixel pairs becomes $K\times(K-1)/2$ and the training matrix is established with $K\times(K-1)/2$ rows and $M+1$ columns. Each row in the training matrix represents the distribution of the pixel difference of the one-pixel pair from the training set. The iterative training process is as follows:

1. All rows of the training matrix are sorted by the discrimination scores so that the pixel pair of the first row has the highest discrimination score.
2. The first row in matrix ***T*** is moved into the result matrix ***R*** to initialize the result matrix.
3. We extract the next row in matrix ***T*** and calculate the correlations between this row and all rows in matrix ***R***.
4. If all correlations are smaller than the predefined threshold C, put this row into result matrix ***R*** and go to stage 3, otherwise, go to stage 3 directly.
5. If the number of rows in the result matrix reaches the predefined N, the iteration stops.
6. If T is empty, stop the iteration.

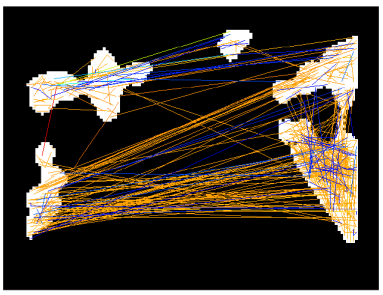


Fig. 4 Extracted pixel pairs

The pixel pairs in the result matrix can be used to compute the binary feature of the test frame. Fig. 4 shows a sample of the extracted pixel pairs. The two endpoints of each line are the pixel pair and the color represents the discrimination score of this pair. Each line represents a pixel pair that contains two pixels on the two endpoints of the line. Pixel pairs with low discrimination scores are drawn in blue while those with high discrimination scores are drawn in red. In the rest of the paper, we entitle our approach as a "Learning-based Binary Feature Approach" or simply say our approach.

IV. EXPERIMENTS

In our experimental work, the extracted key regions, key frames, and binary scene features based on machine learning were used to match the test sequence with the reference sequence. We use the desktop computer as the processing platform.

A. Dataset

The experiment used a light-rail dataset provided by Mass Transit Railway (MTR) in Hong Kong and Nordland dataset published by Norwegian Broadcasting Corporation (NRK) [23]. The dataset of Hong Kong light rail transit (LRT) was collected from route 507, containing 3 sets of video sequences, containing a total of 13,859 frames. Each set of video sequences contains two sequences, which are collected from the same train running on the same path at different times. The two video sequences in the Hong Kong MTR data set were captured by monocular cameras 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 train speed are all different in these two sequences. 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, 10000 frames are used as training and testing data and down-sampled to 640×480. The four sequences keep running at the same speed. Therefore, the frames with the same index number were collected from the same location.

B. Evaluation with Single Frame Scene Recognition

The extracted key frames and key regions in the reference sequences were used to lock tracking with scene recognition. Therefore, the quality of key frames and key regions should be evaluated by the quality of scene recognition. The key frames are expected to lock the tracking by providing a high confidence matching score when the train comes into the same position with the most similar scene appearing again in the current frame.

In the experiment, we compared the following four different methods of scene recognition, including the method based on global feature and the method based on the local key region. We used the HOG feature to evaluate the quality of key regions in the scene matching. The evaluation criterion is the mean error deviation of the matching result and the ground truth.

Firstly, the Global HOG feature means that the whole video frame is described by one HOG descriptor by using this to calculate the difference between two images. Secondly, to retain the relative position information of video frame content, we divided each video frame into 40×40 non-overlapping macroblocks and the HOG features of each block are calculated separately. The current frame and the reference frame were matched with the corresponding macroblocks. Thirdly, to evaluate the performance of ROI in this paper, we only considered the macroblocks within the ROI. At last, the HOG feature in the key region is the proposed method which only matches the current frames and reference frames with HOG features of connected key regions.

After matching the current sequence with the reference sequence, for each current frame, the difference of frame indices between the beat matched reference frame and ground truth reference frame which is called error offset were recorded. The absolute average value of these differences, called average error offset, is used to evaluate the precision of 4 methods. The unit of this average error is a frame. The ideal situation with this error approaching zero means that all current frames are matched with the corresponding ground truth reference frame. Ideally, this offset is close to zero, which means that all matching results are the same as the ground truth. As shown in Table □, the proposed method based on the key region has the lowest average error offset. Macroblock-based HOG feature has the highest time complexity, and the matching time of each frame reaches 62.42s. However, the average error offset increased by 0.16 frames when using ROI to reduce the computation time for scene matching. Comparing with global HOG feature, using the HOG feature in the proposed key region reaches a trade-off between computation time and quality.

TABLE I
COMPUTATIONAL TIME AND MATCHING OFFSET OF MATCHING

Method	Matching offset (frame)	Computational time (Second)
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 Key Region (Proposed)	1.44	3.6058

As described above, the key region of the current frame is determined by the discriminative score and the predefined threshold T_k. T_k is an adaptive threshold in our system because the discriminative scores in each frame are distributed on different scales. So it cannot use the unified absolute threshold. For example, the range of the discriminative score in a key frame has far higher than that in the non-key frame. Therefore, the threshold T_k is indirectly adjusted by the coefficient K. is the product of the average significance score within the frame and the coefficient K. The T_k of a key frame is the product of the average discrimination score and the coefficient K. In the following parts, we will give some discussions about this coefficient K.

Firstly, we test the computation time of scene recognition under different coefficient values. The range of coefficient K is from 0 to 1.40. As shown in Fig. 5, as the coefficient K goes up from 0.75, the computation time of scene recognition decreases rapidly. Therefore, the large coefficient K value makes it possible to apply the approach to single scene recognition in the real-time system. The main reason is that less number of regions are defined as key regions when a larger coefficient K is used. The smaller key regions allow the HOG feature to be generated faster. The basic cell size of the HOG feature in our system is fixed as 10 pixels so that smaller key regions means fewer cells in the description area and the computational complexity of the HOG feature is reduced. The HOG feature used in this system has a fixed basic cell. The use of a smaller key region means that the number of cells in the description region is reduced, thus reducing the computational complexity of the HOG feature.

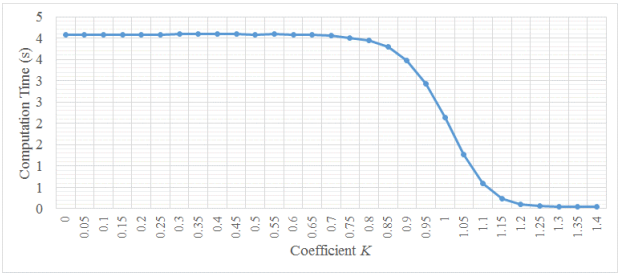


Fig. 5 Computation time of scene recognition with key regions extracted with different coefficients

The percentage of the key region area in the entire video frame under different values of coefficient K shown in Fig. 6 proves our supposition. For each frame, the proportion of the key region area to the whole frame was recorded. The vertical axis is the average proportion of the key region of all frames in the dataset. We can see that the percentage of key regions drops off as the coefficient K increasing. When K equals zero, it means that all pixels in the ROI are regarded as key regions. In this case, the proposed method keeps the system only focused on a quarter of the frame rather than the whole frame.

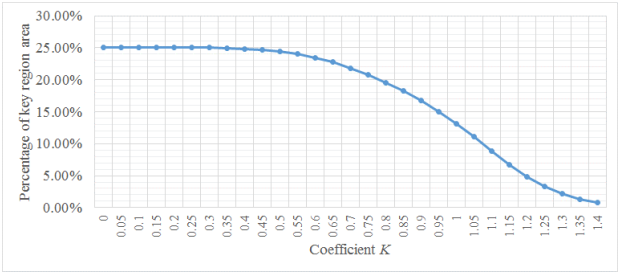


Fig. 6 Percentage of key region area

The trend of computation time and the percentage of key region area reveals the high efficient of a large coefficient. When the coefficient K increases from 0 to 1.4, the computation time of scene recognition drops from 4 seconds to 0.039 second and the percentage of key region drops from 25% to 0.76%. However, a large coefficient K leads frames to

face the risk of the non-key region. Fig. 7 shows that when K becomes larger than 1.1, some of the frames in the dataset will fail to identify the key regions. The percentage of frames without key regions will increase when larger K is used. These frames without any key regions require to match with all pixels in the ROI, which will reduce the efficiency of scene recognition.

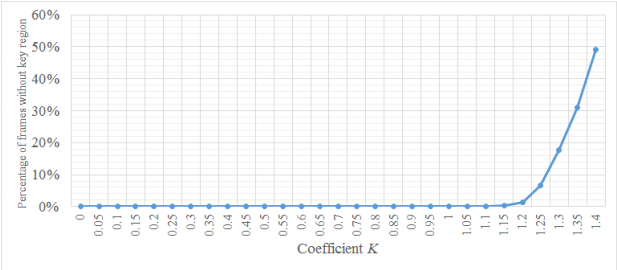


Fig. 7 Percentage of frames without key region

To determine the most appropriate value of coefficient K, we used all available values of K in the dataset for scene recognition and calculated statistics of average error deviation. A set of different cell sizes using the HOG feature was also used to make the result more credible. The set of different cell sizes were obtained by changing the parameter *N* in formula (1). We tested 7 different sizes of cells which range from 5 to 50. Fig. 8 shows the error rate of scene recognition under each kind of parameter selection.

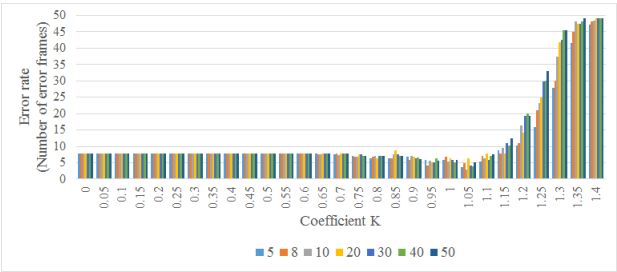


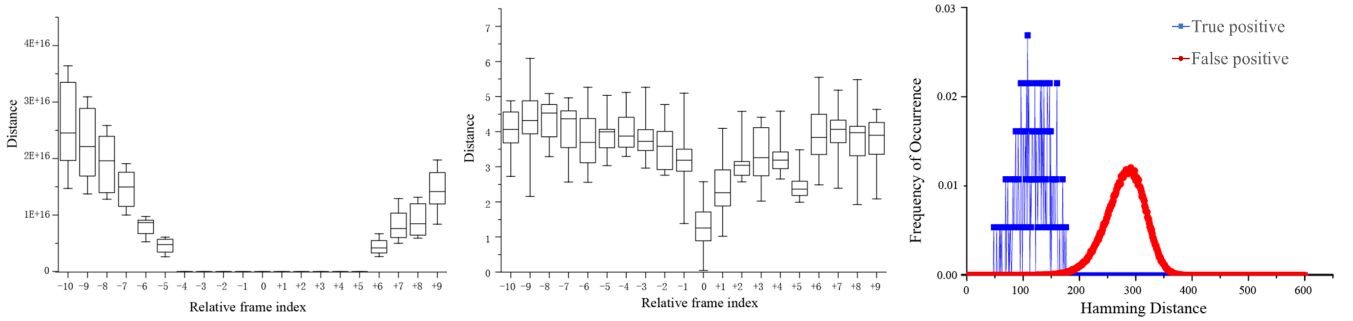
Fig. 8 Error rate of scene recognition with different cell sizes and coefficient *K*

The vertical axis is the error rate of matching, which has the same definition in the first part of this subsection. The bars with different colors present different cell sizes. The error rate indicates that if *K* is too large, it gives a high matching error, such as 1.4. This huge error rate shows that although the most powerful key regions are selected, due to the too few pixels are selected, and visual information is limited and the recognition procedure will be affected by the noise. We can notice that the lowest error rate appears when the coefficient *K* equals 1.05. This result gives us a recommendation value of coefficient *K*, which has the best accuracy of scene recognition and acceptable efficiency. The key regions extracted with a coefficient *K* of 1.05 can be used to do the further training procedure of binary feature extraction.

C. *Evaluation with Multi-frame Scene Tracking*

Multi-frame scene tracking was tested by using the Nordland dataset. The tracking module first uses the SeqSLAM [23] algorithm to match the current sequence with the reference sequence and then obtain a set of candidate reference frames. SeqSLAM, like a scene sequence matching method, is widely used in path-based visual localization algorithms [24-26]. The global features in SeqSLAM were generated by the down-sampling normalized image with a resolution of 32*24 pixels. The similarity of frames was measured by the first norm distance between the current frame and the reference frame in SeqSLAM. Lower distance between two frames means that they have a similar appearance. These results are verified by the binary feature in the matching module. The higher matching score means that the two frames were taken in the same place. The precision is used to evaluate the performance of the tracking result. The true positive (*TP*) is defined as that a positive sample is formed near to the ground truth position within 3 frames. Otherwise, the positive sample is regarded as a false positive (*FP*). The precision can be calculated by *TP/(TP+FP)*.

Fig. 9(a) shows the matching distance distribution of 13 mismatched frames in SeqSLAM. We recorded the distance distribution of 20 frames around the ground truth. The vertical axis is the matching distance and the horizontal axis is related frame index in the temporal domain. About 10 frames before and after the ground truth have a similar matching distance and it makes the SeqSLAM tracking module can hardly identify the best-matched frame. When these frames were



(a) Distribution of matching distance of SeqSLAM (b) Distribution of matching score of binary feature (c) Distribution of matching distance frequency based

Fig. 9 This method is compared with the results of SeqSLAM in high frame rate scene matching

verified by the binary feature, the matching scores always have a peak on the ground truth frame, as shown in Fig. 9(b). This indicates that the Learning based Binary feature Approach only gives low matching distances in ground truth locations and allows the matching module to provide an accurate result with higher confidence.

Furthermore, to verify the performance of the binary feature proposed in this paper, we also calculate the distribution of the binary matching scores of the true positive and false positive in SeqSLAM with the whole dataset. Fig. 9 (c) gives a proof for our assumption. The blue curve is the frequency of the true positives and the red curve is that of false positives. These two curves are separated significantly with a Hamming distance of 75. Therefore, our approach can distinguish similar frames and improve the precision of the SeqSLAM tracker.

The Table □ shows the precision, matching offset and matching time of the two tracking algorithms. The precision is improved to 99.36% by using the binary feature approach proposed in this paper. The matching offset decreased by 36.07% without significantly increasing the scene matching time. These results show that the Learning based Binary feature Approach can provide more significant visual information for scene recognition. Thus the approach can obtain more accurate matching, while the global feature in SeqSLAM can only provide a rough matching result.

TABLE □
PRECISION AND COMPUTATION TIME OF SCENE TRACKING

	SeqSLAM	Proposed method	Δ (%)
Precision	89.56%	99.36%	+9.80
Matching offset (frame)	1.3652	0.8728	-36.07
Time (ms)	53.23	54.82	+2.99

D. Evaluation for Keyframe-based Retrieval Mechanism

To verify the necessity of keyframe-based retrieval mechanism, we compare the proposed scene tracking method with global tracking and local tracking. Fig. 10 shows the tracking route of the three scene matching methods, where the black curve represents the ground-truth path and the white curve represents the matching result. All the tracking methods make use of the key region detection and the learning-based binary features extraction proposed in this paper.

The global tracking method calculates the similarity between the current frame and all reference frames, to retrieve the reference frame with the globally optimal matching in the whole reference sequence. The disadvantage of this method is that it is easily affected by other similar scenes in the path. As shown in Fig. 10 (a), due to the problem of vehicle occlusion, it directly jumps to other similar scenes. When the scene is severely disturbed, the global tracking will take similar scenes of other road sections as localization results.

The local tracking method only matches the current frame and the reference frame within the neighbourhood of the last matching result, which makes full use of the spatial constraints between scenes. However, its disadvantage is that when the empirical tracking results are not good, it will directly affect the results and cause cumulative errors. As shown in Fig. 10 (b), the failure is caused by the cumulative tracking error which lead the route to deviating from the correct range after several consecutive wrong matches.

In Fig. 10 (c), the scene matching result (white curve) almost coincides with the ground truth (black curve) because the keyframe-based retrieval mechanism guarantees the correctness of the scene tracking result. Though, at the red dotted line, the trajectory score of the reality tracking route is not globally optimal. The experimental results show that the scene matching method based on key frame retrieval mechanism can reduce the matching offset to an acceptable range and avoid the cumulative error, which can effectively reduce the interference of scene matching caused by some extreme conditions in Hong Kong LRT dataset.

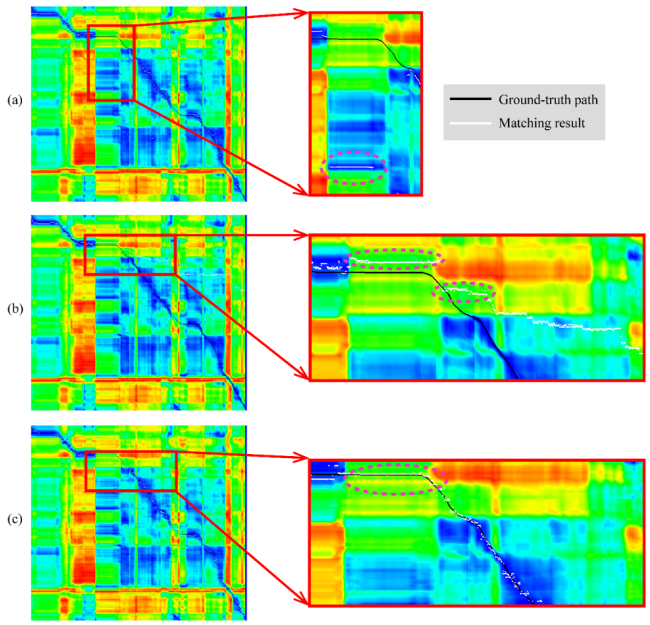


Fig. 10 Tracking routes of the three scene matching methods. (a) Global tracking method. (b) Local tracking method. (c) The proposed method.

V. CONCLUSIONS

The LRT localization system based on monocular camera has the advantages of simple acquisition equipment and good application prospects. The main difficulty of the system is the high computational complexity and low accuracy. Therefore, we proposed a light rail localization method based on key regions and unsupervised learning. The extraction of key regions not only improves the accuracy of single scene recognition but also reduces the computational time cost. Secondly, we design a new saliency measurement standard to realize binary features extraction in key regions. Finally, this paper implements a real-time localization system for light rail based on a single viewpoint video. The experimental results

show that the Learning based Binary Feature Approach raises the matching accuracy and the computation time is extremely low, which is suitable for real-time applications.

ACKNOWLEDGMENT

This paper is supported by the Project for the National Natural Science Foundation of China under Grant No. 61672064, the Beijing Natural Science Foundation under Grant No. 4172001 and KZ201610005007, and Beijing Laboratory of Advanced Information Networks under Grant No. PXM2019_014204_500029.

REFERENCES

[1] N. Piasco, D. Sidibé, C. Demonceaux, V. Gouet-Brunet, "A survey on Visual-Based Localization: On the benefit of heterogeneous data," in *Pattern Recognition*, vol. 74, pp. 90-109, February 2018.

[2] F. Amorós, L. Payá, J. M. Marín, O. Reinoso, "Trajectory estimation and optimization through loop closure detection, using omnidirectional imaging and global-appearance descriptors," in *Expert Systems with Applications*, vol. 102, pp. 273-290, July 2018.

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

[4] 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.

[5] W. Maddern, S. Vidas, "Towards robust night and day place recognition using visible and thermal imaging," unpublished.

[6] J. Zhang, S. Singh, "Laser-visual-inertial odometry and mapping with high robustness and low drift" in *Journal of Field Robotics*, vol. 35, no. 8, pp. 1242-1264, December 2018.

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

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

[9] D.G. Lowe, "Distinctive image features from scale-invariant key points." in *International Journal of Computer Vision*, vol 60, no. 2, pp: 91-110, November 2004.

[10] H. Bay., A. Ess, T. Tuytelaars, L. V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol 110, no. 3, Pages 346-359, June 2008.

[11] E. Rosten, R. Porter and T. Drummond, "Faster and Better: A Machine Learning Approach to Corner Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105-119, Jan. 2010.

[12] M. Calonder, V. Lepetit, C. Strecha, P. Fua, "Brief: Binary robust independent elementary features," *European Conference on Computer Vision*, Springer Berlin Heidelberg, vol. 5, pp: 778-792, September 2010.

[13] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *2011 International Conference on Computer Vision*, Barcelona, pp. 2564-2571, 2011.

[14] C. McManus, B.Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," *Robot. Sci. Syst. Conf.*, Berkeley, CA, USA, July 2014.

[15] F. Han, X. Yang, Y. M. Deng, M. Rentschler, D. J. Yang, H. Zhang, "SRAL: Shared representative appearance learning for long-term visual place recognition," in *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1172-1179, April 2017.

[16] N. Carlevaris-Bianco, R. M. Eustice, "Learning visual feature descriptors for dynamic lighting conditions," In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicago, USA, pp. 2769-2776, September 2014.

[17] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp: 1097-1105, 2012.

[18] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, M. Milford, "Place recognition with ConvNet land marks: Viewpoint-robust, condition-robust, training-free," in *Robot. Sci. Syst. Conf.*, Rome, Italy, July 2015.

[19] C. Zitnick and P. Doll´ ar, "Edge Boxes: Locating object proposals from edges," in *Proc. 13th Eur. Conf. Comput. Vis.*, pp. 391–405, 2014.

[20] R. Arroyo, F. P. Alcantarilla, L. M. Bergasa, E. Romera, "Fusion and binarization of CNN features for robust topological localization across seasons." In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4656-4663, 2016.

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

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, vol. 1, pp. 886-893, 2005.

[23] 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, USA, pp. 1643-1649, 2012.

[24] G. Bresson, Z. Alsayed, Y. Li, S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," in *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194-220, September 2017.

[25] P. Kim, B. Coltin, O. Alexandrov, H. J. Kim, "Robust visual localization in changing lighting conditions," In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5447-5452, May 2017.

[26] D. Bai, C. Wang, B. Zhang, X. Yi, X. Yang, "Sequence searching with CNN features for robust and fast visual place recognition," in *Computers & Graphics*, vol. 70, pp.270-280, February 2018.