Dynamic Environments Localization via Dimensions Reduction of Deep Learning Features

Hui Zhang[1(✉)], Xiangwei Wang[1], Xiaoguo Du[1], Ming Liu[2], and Qijun Chen[1]

[1] RAI-LAB, Tongji University, Shanghai, China
huizhang629@gmail.com
[2] RAM-LAB, Robotics Insititute, HKUST, Hongkong, China

Abstract. How to autonomous locate a robot quickly and accurately in dynamic environments is a primary problem for reliable robot navigation. Monocular visual localization combined with deep learning has gained incredible results. However, the features extracted from deep learning are of huge dimensions and the matching algorithm is complex. How to reduce dimensions with precise localization is one of the difficulties. This paper presents a novel approach for robot localization by training in dynamic environments in a large scale. We extracted features from AlexNet and reduced dimensions of features with IPCA, and what's more, we reduced ambiguities with kernel method, normalization and morphology processing to matching matrix. Finally, we detected best matching sequence online in dynamic environments across seasons. Our localization algorithm can locate robots quickly with high accuracy.

1 Introduction

Where am I? It's the primary problem in reliable robot navigation to locate quickly and accurately in changing environments. Such changes come from many sources including dynamic objects, varying weather and season shifts. An intelligent robot must be equipped with the ability to adapt to these changes. It doesn't conform to reality that automatic driving cars can only run in the trained scenes. So it's essential to express the scene images without the influence of substantial changes. Over the past few years, various types of features have been investigate for localization [2,7,20,27]. Image descriptors can be divided into feature_based and holistic image descriptor. Features_based descriptors play an important role in Computer Vision. Up to now, several hand-crafted features have gained some success [3,16,23,30]. However, the robots often fail to locate themselves in dynamic environments with these hand-crafted feature descriptors.

Holistic images descriptor express one image according to invariant features. Deep-learning has dramatically changed the overnight. It greatly boosted the development of visual perception, object detection and speech recognition [29]. Recent results indicated that the generic descriptors extracted from the convolutional neural networks are very powerful [26].

M. Liu et al. (Eds.): ICVS 2017, LNCS 10528, pp. 239–253, 2017.
DOI: 10.1007/978-3-319-68345-4_22

In 2012, CNNs got incredible accuracy on the AlexNet Large Scale Visual Recognition Challenge (ILSVRC) [10]. It suggested that features extracted from CNNs significantly outperformed hand-crafted features on classification. They trained a large CNN named AlexNet with 1.2 million labeled images. Because the images are classified according to the features extracted from AlexNet, we can also locate robots based on these features. [8] indicated that features from mid-layer of CNNs can remove dataset bias more efficiently. [28] compared the performance of features from different layers. Their results showed that features from the middle layers in the ConvNet hierarchy exhibited robustness against appearance changes induced by the time of day, seasons, or weather conditions. Features from Conv3 layer performs reasonably well in terms of appearance changes.

Nevertheless, the main obstacle of CNNs features is expensive computational costs and memory resources, which is a big challenge for real-time performance. [1] compressed the redundant data of CNN features into a tractable number of bits. The final descriptor is reduced by applying simple compression and binarization techniques for fast matching using the Hamming distance. It's necessary to reduce the dimensions of these vectors. Compression means losing some information. However, we can keep important relationship among data as much as possible. We realize this purpose through Incremental PCA (Principal Component Analysis) that used widely in data analysis [31].

In this paper, we present a novel algorithm to locate a robot in dynamic environments across seasons. The main contributions of this paper are: (1) We proposed a novel localization system in dynamic environments via dimensions reduction of deep learning features. (2) We reduced the dimensions of features extracted from AlexNet. It can not only quicken computing speed but also reduce confusing matching from datasets. (3) Instead of complex data association graph, we found best matching sequence online with morphology processing to matching matrix.

2 Related Work

2.1 Feature Extraction

It's a big challenge to express a scene that changing significantly, shown in Fig. 1. The recent literature proposed a variety of approaches to address the challenge of this field [5,6,18–22]. As we all know, CNNs got incredible accuracy on the AlexNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 [10]. [8–10,25,26] proved that ConvNets have been demonstrated outperforms traditional hand-crafted features [3,3,16,23]. This network consists of five convolutional layers followed by three fully connected layers and a soft-max layer. It was pre-trained with 1.2 million labeled images. The images are classified according to the features extracted from AlexNet. The output of each individual layer can be used as a global image descriptor. We can also match images based on these features and then locate robots. [8] indicated that features from mid-layer of CNNs can remove dataset bias more efficiently. [28] compared the performance

of different layers features. Their results showed that features from the middle layers in the ConvNet hierarchy exhibit robustness against appearance changes induced by the time of day, seasons, or weather conditions. Features from Conv3 layer performed reasonably well in terms of extream appearance changes. The vector dimensions of different layers in AlexNet ConvNets are listed in Table 1.

(a) Rainy night (b) Rainy daytime (c) Shadows of trees and buildings and others

(d) Moving objects (e) Dusk (f) Light outsides

Fig. 1. Dynamic environments including dynamic objects, varying weather and season shifts.

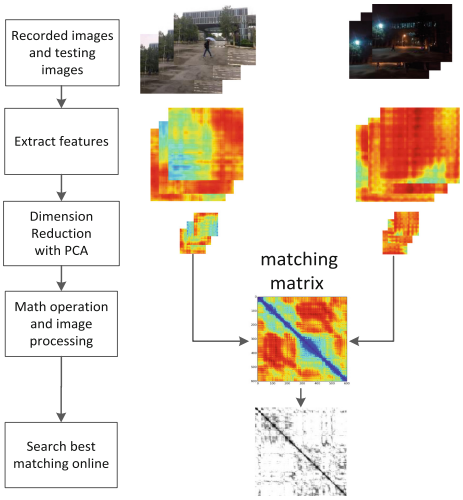


Fig. 2. Outline illustration of dynamic environments localization via dimensions reduction of deep learning features. Features of training images and online images are all extracted from AlexNet.

[28] proved that features from Conv3 layer performed reasonably well in terms of extream appearance changes. Besides, [28] also pointed that fc6 and fc7 outperform the rest layer in terms of viewpoint changes. However, fc6 and fc7 fail completely when appearance changes.

The dimensions of Conv3 are 64896, which means that one image is shown as a 64896 dimensions vector. Online localization will receive images from camera continuously. There is no doubt that a large number of vectors math operation is time-consuming. The different features contained in DCNN are initially returned in a float format. With the aim of facilitating a subsequent binarization, [1] cast these features into a normalized 8-bit integer format. Then a matching matrix is computed by matching all the binary features using the Hamming distance. Their results showed that compression of features can reduce the 99.59% redundancy of their descriptors, while precision is only decreased in about 2%. Besides, their binarization of features allowed using the Hamming distance, that also represented a speedup to match locations.

2.2 Image Matching

Image matching is another challenge after features extraction. By the way, image matching means place recognition in robot localization domain. There is no doubt that the robot's knowledge of the world must be stored as a map, to which the current observation is compared. [17] pointed out that the map framework differs depending on visual sensors and what type of place recognition is being performed. They can be divided into pure image retrieval, topological maps, and topological-metric maps. Pure image retrieval only stores appearance information about each place in the environment with no associated position information, just like Chow-Liu tree used in FAB-MAP [7]. FAB-MAP [7] described a probabilistic approach to the problem of matching images and map augment. They used vector-based descriptors like SURF jointly with bags-of-words. This paper learned a generative model of place appearance. They constructed a Chow-Liu tree [4] to capture the co-occurrence statistics of the visual words. Chow-Liu tree is composed of nodes and edges. Mutual information between variables is shown by the thickness of tree's edges. Each node in the graph corresponds to a bag-of-words representation that converted from input sensory data. FAB-MAP was successful in detecting large portions of loop closures in challenging outdoor environments. But results of [21] show that in datasets over seasons only a few correct matches are found by OpenFABMAP2 due to that the hand-crafted feature descriptors are not repeatable. Paper [21] formulated image matching as a minimum cost flow problem in a data association graph to effectively exploit sequence information. They locate vehicle through Minimum Cost Flow. Their method worked well in dynamic scenes. [12] presented a Markov semi-supervised clustering approach and its application in topological map extraction. As for incremental mapping, slam, and navigation tasks, the approach can be adapted accordingly.

SeqSLAM [20] framed the image recognition problem as one of finding all the templates within local neighborhoods that are the best matching for the

current image. It is easy to implement. However, the algorithm of [20] can easily be affected by robot speed. This constraint limits the applications for long-time localization. [24] proved that place recognition performance improves if only the most informative features from each image are used. [14] described a lightweight novel scene recognition method using an adaptive descriptor, which is based on color features and geometric information. [13] presented a scene recognition approach with omnidirectional vision for topological map using lightweight adaptive descriptors. [11] improved place recognition with a reduced feature set. [15] proposed a generic framework for recognition and clustering problem using a non-parametric Dirichlet hierarchical model, named DP-Fusion.

The paper proceeds as follows. In Sect. 3, we describe details of our methodology. Section 4 gives out the experiment results of online localization in dynamic environments on Norland datasets. In Sect. 5, we have a discussion about the results and future work.

3 Approach and Methodology

In this paper, we contribute a new proposal that exploits the advantages of powerful feature representations via CNNs in order to perform a robust vision-based localization across the seasons of the year, as introduced in the graphical explanation of our approach given in Fig. 2. Our work proceeds as follows.

(1) Extract features from Conv3 of AlexNet. Consider dimensions reduction via IPCA.
(2) Vectors of online images will match with datasets vectors one by one through cosine distance. Normalize matching matrix through kernel method to reduce ambiguities caused by confusing datasets. Save matching matrix as a gray image.
(3) Image processing to the gray matching image including image binarization.
(4) Set parameters and find best matching sequence online through RANSAC (random sample consensus).

3.1 Algorithm Framework

The algorithm framework of our method is described in Algorithm 1. About the map framework, we used pure image retrieval but the datasets were stored in order according to the images' incoming time. If so, we can not only ensure accuracy but also compute efficiently. We chose features from Conv3 of AlexNet as our holistic image descriptor. The dimensions of Conv3 are 64896, which means that one image is shown as a 64896 dimensions vector $\mathbf{f}$. We build visual map $\{[\mathbf{f}, \mathbf{l}]\}_{i=1}^{n}$ with location of each image. So the current image sequences are expressed as $\{\mathbf{I}\}_{j=t-m+1}^{t}$. High-dimensional vectors result in time-consuming. We consider dimensions reduction via IPCA. Although image descriptors are somewhat losing information, it reduces the ambiguous matching causing from the confusing datasets like sky, ground and trees. Vectors of online images will

Algorithm 1. Algorithm: visual localization

Input: Visual Map $\{[\mathbf{f}, \mathbf{l}]\}_{i=1}^{n}$, where $\mathbf{f}$ is the feature vector of image on location extracted from AlexNet; $\mathbf{l}$ is location of corresponding image; n is the size of visual map; Current image sequences $\{\mathbf{I}\}_{j=t-m+1}^{t}$, where m is the sequence size; last robot location $\hat{l}_{t-1}$

Initialize: $\hat{l}_t = \hat{l}_{t-1}$ **Output**: Robot current location l_t

for $t = 2$ *to* n **do**

- Calculate $\{\hat{\mathbf{f}}\}_{j=t-m+1}^{t}$ the feature of $\{\mathbf{I}\}_{j=t-m+1}^{t}$
- Calculate matching matrix $\mathbf{M}$ and $\mathbf{M}_{ij} = \mathcal{F}(\mathbf{f}_i, \hat{\mathbf{f}}_j)$
- Kernel process to every element of the matching matrix $\hat{\mathbf{m}}_{ij} = e^{1-\cos \mathbf{m}_{ij}}$
- Normalize the matching matrix $\mathbf{M}_{ij} = \frac{255(\mathbf{M}_{ij} - \mathbf{M}_{min})}{\mathbf{M}_{max} - \mathbf{M}_{min}}$ take $\mathbf{M}$ as a gray image $\mathbf{I}_g$
- Change $\mathbf{I}_g$ to binary image $\mathbf{I}_b$ using suitable thresholding.
- Deal $\mathbf{I}_b$ with morphology method and get image $\mathbf{I}_m$
- Using RANSAC method to find the best matching line $y = kx + b$ on $\mathbf{I}_m$
- The current image's best matching feature in the visual map is $\mathbf{f}_{km+b}$
- Set $\hat{l}_t = l_{km+b}$

end

return $\hat{l}_t$

be compared with datasets vectors one by one through cosine distance. We then get matching matrix $\mathbf{S}$ whose elements float in range (0, 1]. Normalize matching matrix through kernel method to reduce ambiguities caused by confusing datasets that match against most of the online images. Then it is converted to a binary gray image by a suitable thresholding. We tried to adjust parameters and then find the best matching sequence online through RANSAC. The current image's best matching feature in matching matrix is $\mathbf{f}_{km+b}$. Then the current image's best matching image in the visual map is $\mathbf{l}_{km+b}$.

3.2 Feature Extraction from Deep Learning

We extracted features from Conv3 of AlexNet as our image holistic descriptor provided by Caffe. The dimensions of Conv3 are 64896, which means that one image is expressed by a 64896 dimensions vector. The vector dimensions of different layer in AlexNet ConvNets are listed in Table 1 [10]. [28] gave us the conclusion that the layers higher in the hierarchy are more semantically meaningful but therefore lose their ability to discriminate between individual places within the same semantic type of scene. It's important to decide which layer we use. Features from Conv3 layer performs reasonably well in terms of extream appearance changes.

3.3 Dimensions Reduction

We tested on Norland datasets to determine how many dimensions fit best for time consuming and accuracy. We chose 300 images sequence in the spring season

Table 1. Dimensions of different layer of AlexNet

Layer	Dimensions	Layer	Dimensions
Conv1	$96 \times 55 \times 55$	Conv4	$384 \times 13 \times 13$
pool1	$96 \times 27 \times 27$	Conv5	$256 \times 13 \times 13$
Conv2	$256 \times 27 \times 27$	fc6	$4096 \times 1 \times 1$
pool2	$256 \times 13 \times 13$	fc7	$4096 \times 1 \times 1$
Conv3	$384 \times 13 \times 13$	fc8	$1000 \times 1 \times 1$

Table 2. Relationship between percent of main information and n_components

n_components	Information ratio	n_components	Information ratio
316	99%	51	93%
187	98%	44	92%
136	97%	38	91%
99	96%	33	90%
76	95%	29	89%
62	94%	25	88%

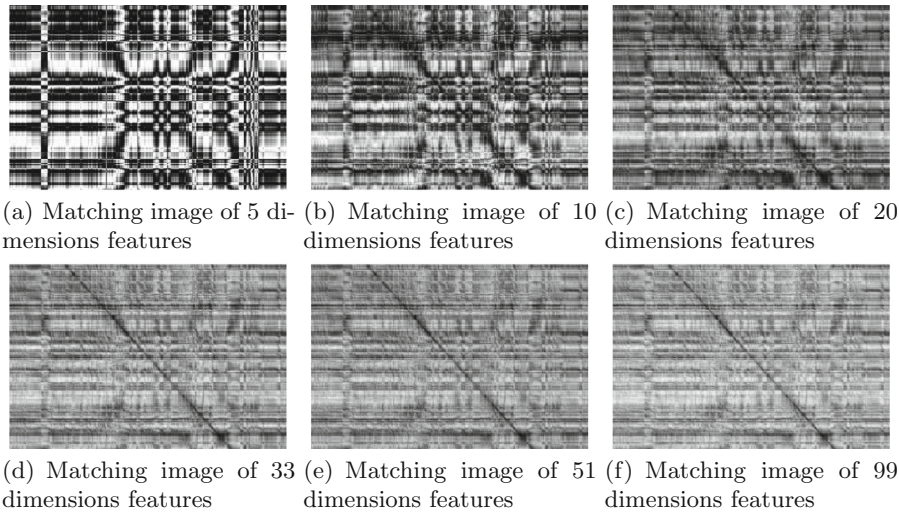
as recorded and 500 images sequence in fall as a test. We used Incremental PCA in scikit-learn for a large number of images matching. IPCA is one of the essential high-dimensional data analysis. IPCA transforms high-dimensional data to low dimensions through the linear transformation. The dimensions of the different layers of AlexNet are shown in Table 1. It is easy to understand that more dimensions we keep more information we will attain, but also time-consuming. So the primary task is to determine how many dimensions we keep for each vector.

The relationships between the parameter n_components and main information Ratio are listed in Table 2. In general, we had better keep at least 90% main information ratio in case of influence on accuracy. We also compared matching result among different dimensions. The comparison results are shown in Fig. 3. The best matching line cannot be detected with less than 20 dimensions. 33 dimensions is clear enough and also save computation consuming. In short, we chose 33 dimensions vectors as image descriptors.

3.4 Kernel Transform and Normalization of Matching Matrix

Our task is to find the best matching line precisely. We have to use math transform to make this line clearer. We choose kernel method including inverse the elements of matching matrix and exponentiation. The reasons for choosing this method are listed as follows.

(1) Cosine distance between 2 images cannot stand for the positive proportion between the similarity and the matching matrix elements.



(a) Matching image of 5 dimensions features (b) Matching image of 10 dimensions features (c) Matching image of 20 dimensions features

(d) Matching image of 33 dimensions features (e) Matching image of 51 dimensions features (f) Matching image of 99 dimensions features

Fig. 3. Comparision matching image among different dimensions including 5, 10, 20, 33, 51, 99 dimensions. The best matching line turns clear with dimensions increasing.

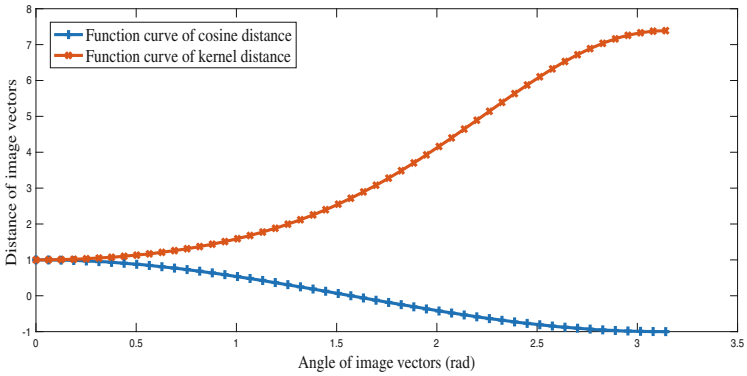


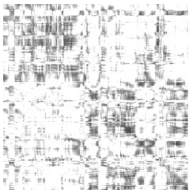
Fig. 4. Function curves of cosine and kernel method distance.

(2) Kernel method will widen the distance between false negative and true positive places.

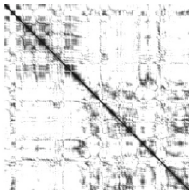
Figure 4 is function curves comparison computed from cosine distance, shown in Eq. (1), and kernel method distance, shown in Eq. (2). The blue line stands for the cosine distance of two image vectors. The brown line stands for the kernel method distance. We can see that kernel method can augment the difference between totally different and similar places. The color of the best matching line appeared as black and the different places appeared as white, shown in Fig. 5. What's more, normalize matching matrix through kernel method reduced ambiguities that caused by confusing datasets that match against most of the

online images. Save matching matrix as a gray image for the following processing including morphology transformation and binarization.

What's more, we normalized the matching matrix with the range of 0 to 255 with Eq. (3). It turned evident after kernel method. It's of great help for the morphology processing and visualization.



(a) Matching matrix of cosine distance



(b) matching matrix after kernel method

Fig. 5. We chose a sequence of 3000 images in spring Norland datasets as trained features and 3000 images in winter Norland datasets as online images. (a) Cosine distance matching matrix. (b) Kernel method distance matching matrix.

We tested kernel method on spring and winter seasons in Norland datasets. There are 3000 spring images and 3000 winter images captured in the same place. Besides, the beginning of two images sequence is the same image. Thus, one line appears on the diagonal for it's the best matching sequence. The matching result is shown in Fig. 5. We matched online images with recorded datasets images one by one through cosine distance with $cos <\mathbf{f}_i, \mathbf{f}_j>$. However, the matching image shown in Fig. 5(a) appeared confusion between terrible matching and perfect matching. However, the diagonal line becomes evident through kernel method of Eq. (2) and normalization method of Eq. (3). The matching image is shown in Fig. 5(b). At last but not the least, save the matching matrix as a gray image, which will be converted to binary one by suitable threshold.

$$\cos <\mathbf{f}_i, \mathbf{f}_j> = \frac{\sum_{i=1}^{33} a_i b_i}{\sum_{j=1}^{33} a_j^2 \sum_{k=1}^{33} b_k^2} \tag{1}$$

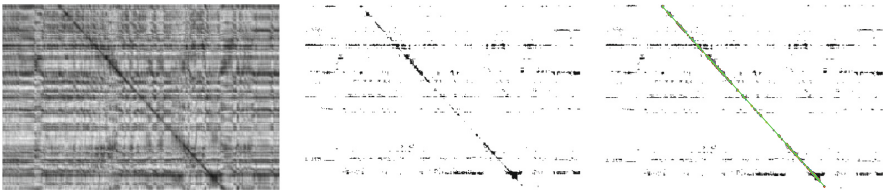
$\mathbf{f}_i = \{\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_{33}}\}$, $i \in D$, D is set of datasets images, $\mathbf{f}_j = \{\mathbf{b_1}\ \mathbf{b_2}\ \ldots\ \mathbf{b_{33}}\}$, $j \in O$, O is set of online images

$$\hat{\mathbf{m}}_{ij} = e^{1-\cos \mathbf{m}_{ij}} \tag{2}$$

$$\mathbf{M}_{ij} = \frac{255\,(\mathbf{M}_{ij} - \mathbf{M}_{min})}{\mathbf{M}_{max} - \mathbf{M}_{min}} \tag{3}$$

4 Experiments

Our experiments are designed to show the capabilities of our method with reduced features and image processing. Our approach is able to (i) localization in scenes across seasons ignoring dynamic objects, varying weather and season shifts. (ii) save time and computation consuming. We perform the evaluations on public available Norland datasets used in SeqSLAM [20]. The gray images were captured in 1 frame every second and the size have been cropped into 64×32. If our approach still works in such unclear and tiny images, then it can save a lot of time and computation consuming. Examples of matching images are shown in Fig. 3.



(a) Matching matrix after kernel method and normalization

(b) Binarization image

(c) Best matching line in matching image

Fig. 6. We chose a sequence of 300 images in fall season trained as map and locate online in spring season. (a) Matching image after kernel method and normalization. (b) Binarization image of (a) with suitable thresholding. (c) Detecting line with RANSAC algorithm and the green line is just the best matching line. (Color figure online)

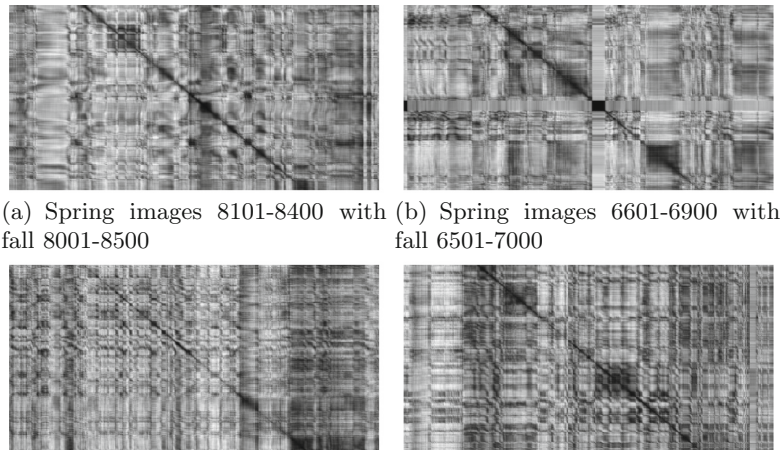
We can see that in Fig. 5(b) the best matching line became obvious. Our task is to find its mathematical model to find the corresponding index in datasets. We decided to use classical RANSAC algorithm.

4.1 Online Search in Dynamic Environments

In Fig. 6, we chose a sequence of 300 images in fall season trained as the map and locate online in the spring season. We can see that the features extracted from Conv3 of AlexNet didn't affect the matching result. On the opposite, reduce the influence of background information, shown in Fig. 6(a). Figure 6(b) is the binarization result of the matching image. You see that most of the interfere information has been wiped off. The capacity of restraining distractor has more important effect during robots localization. We can see that in Fig. 6(c) the green line is just the best matching in this period. The current image's best matching feature in matching matrix is $\mathbf{f}_{km+b}$. Then the current image's best matching image in the visual map is $\mathbf{l}_{km+b}$.

In Fig. 8, we plot 3 lines to assess the error of our approach. The blue line stands for the index of ground truth. Red line means the index of matched

images with our approach. The yellow one is index error between ground truth and matching images. The search index in the range [1872, 2026] in x coordinate axes cannot be updated. We will discuss this problem in Sect. 5 (Fig. 7).



(a) Spring images 8101-8400 with fall 8001-8500

(b) Spring images 6601-6900 with fall 6501-7000

(c) Spring images 9301-9600 with fall 9201-9700

(d) Spring images 9601-9900 with fall 9501-10000

Fig. 7. Examples of some matching images. The number, take '8101–8400' for example, means index of images sequence.

4.2 Results

Our paper present a novel and time-consuming algorithm to locate a robot in dynamic environments across seasons. It's a rapid localization system. We extracted features from Conv3 of AlexNet and it did outperform hand-crafted features in robots localization domain. Dimensions reduction via IPCA is a novel try. Each layer of AlexNet develops advantages in the different domain. It's proved that Conv3 is the best choice for robots localization. Luckily it helped a lot to quicken computing speed and reduce confusing matching from datasets caused by images match against most of the online images. We compared vectors of online images with datasets vectors one by one through kernel method distance. This process widens the difference between similar and totally different places. What's more, image processing to the gray matching image, including converting to binary one by suitable thresholding, turned complex data association graph into simple image processing. As for sequence matching, we used classical RANSAC algorithm to find the best matching line. Our experiments results show that dimensions reduction is a great idea to quicken computing speed and reduce confusing matching. And our algorithm is robust to season shifts, dynamic environments, changing weathers and so on. Examples of some matched images are shown in Fig. 10.

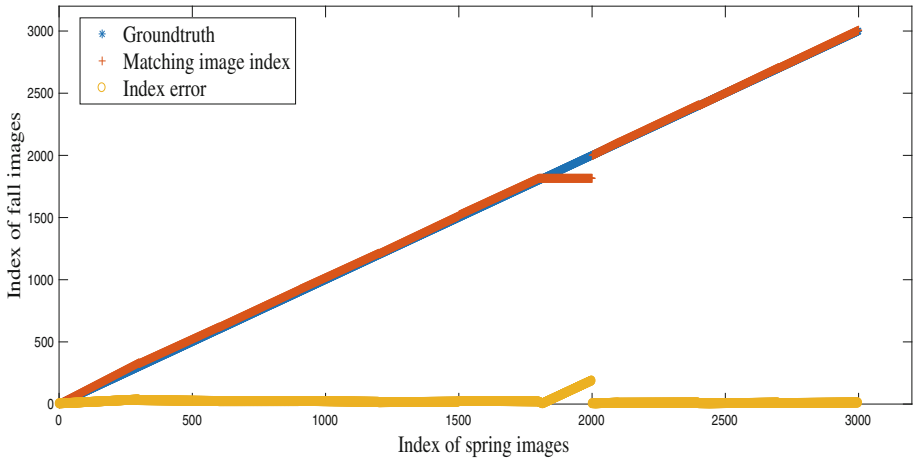
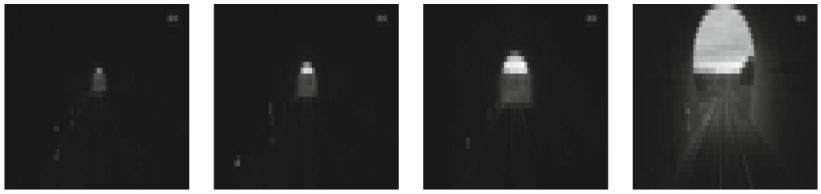
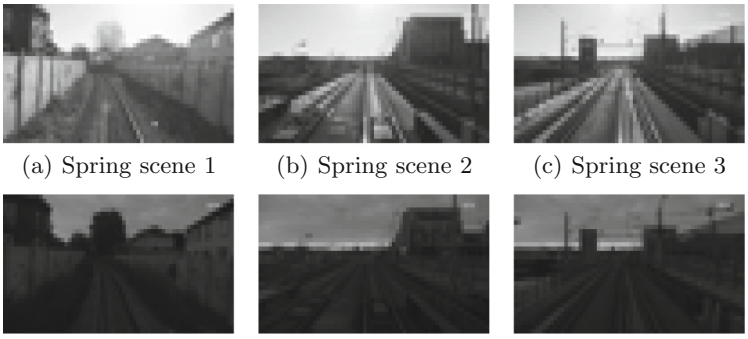


Fig. 8. Experiments result tested on 3000 images. The function of ground truth index line is y = x, shown in blue line. The brown line is matched images index with sequence of 300 images each time. The yellow line is error between real position and matched. (Color figure online)



(a) Images-02204 in spring datasets (b) Images-02205 in spring datasets (c) Images-02206 in spring datasets (d) Images-02207 in spring datasets

Fig. 9. Captured images when the train went through a tunnel in Norland datasets.



(a) Spring scene 1 (b) Spring scene 2 (c) Spring scene 3

(d) Fall scene 1 (e) Fall scene 2 (f) Fall scene 3

Fig. 10. Examples of some matched images.

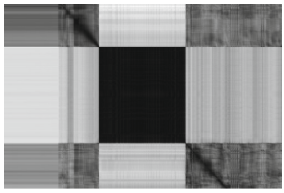


Fig. 11. Matching image that covers totally dark images in spring image 1872 to 2026.

5 Discussion and Future Works

The limitation of our system is from the image capture equipment. The images are hard to express in totally dark surroundings. The matching matrix that covers dark images sequence is shown in Fig. 11. Actually in Fig. 8 there is no matching line for images 1872 to 2026, so we cannot detect the matching line at all. Examples of dark images are shown in Fig. 9. The matching image is shown as a black block. We will consider about adding assist of the laser. Besides, the concrete relationship between features dimensions and the localization accuracy will be studied. We want to find out the most suitable dimensions of CNNs features to ensure precision and operation speed. It needs iterative testing. Besides, we will train a generic holistic image descriptor ignoring the influence of season shift, weather changes, dynamic environments and so on. However, it needs a large number of images captured over years to train CNN.

Acknowledgment. This research is a cooperation work between RAM-LAB of HKUST and RAI-LAB of Tongji University. Our work is supported by National Natural Science Foundation (61573260), Natural Science Foundation of Shanghai (16JC1401200); Shenzhen Science, Technology and Innovation Commission (SZSTI) (JCYJ20160428154842603 and JCYJ20160401100022706); partially supported by the HKUST Project (IGN16EG12).

References

1. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Romera, E.: 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. IEEE (2016)
2. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Yebes, J.J., Bronte, S.: Fast and effective visual place recognition using binary codes and disparity information. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), pp. 3089–3094. IEEE (2014)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). doi:10.1007/11744023_32
4. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**(3), 462–467 (1968)

5. Churchill, W., Newman, P.: Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 4525–4532. IEEE (2012)
6. Corke, P., Paul, R., Churchill, W., Newman, P.: Dealing with shadows: capturing intrinsic scene appearance for image-based outdoor localisation. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2085–2092. IEEE (2013)
7. Cummins, M., Newman, P.: FAB-MAP: probabilistic localization and mapping in the space of appearance. Int. J. Robot. Res. **27**(6), 647–665 (2008)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DECAF: a deep convolutional activation feature for generic visual recognition. In: ICML, vol. 32, pp. 647–655 (2014)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Li, F., Kosecka, J.: Probabilistic location recognition using reduced feature set. In: Proceedings of 2006 IEEE International Conference on Robotics and Automation, ICRA 2006, pp. 3405–3410. IEEE (2006)
12. Liu, M., Colas, F., Pomerleau, F., Siegwart, R.: A Markov semi-supervised clustering approach and its application in topological map extraction. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4743–4748. IEEE (2012)
13. Liu, M., Scaramuzza, D., Pradalier, C., Siegwart, R., Chen, Q.: Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 116–121. IEEE (2009)
14. Liu, M., Siegwart, R.: Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera. IEEE Trans. Robot. **30**(2), 310–324 (2014)
15. Liu, M., Wang, L., Siegwart, R.: DP-fusion: a generic framework for online multi sensor recognition. In: 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 7–12. IEEE (2012)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
17. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: a survey. IEEE Trans. Robot. **32**(1), 1–19 (2016)
18. Lowry, S.M., Milford, M.J., Wyeth, G.F.: Transforming morning to afternoon using linear regression techniques. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 3950–3955. IEEE (2014)
19. McManus, C., Upcroft, B., Newman, P.: Learning place-dependant features for long-term vision-based localisation. Auton. Rob. **39**(3), 363–387 (2015)
20. Milford, M.J., Wyeth, G.F.: SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 1643–1649. IEEE (2012)
21. Naseer, T., Spinello, L., Burgard, W., Stachniss, C.: Robust visual robot localization across seasons using network flows. In: AAAI, pp. 2564–2570 (2014)

22. Neubert, P., Sünderhauf, N., Protzel, P.: Superpixel-based appearance change prediction for long-term navigation across seasons. Robot. Auton. Syst. **69**, 15–27 (2015)
23. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571. IEEE (2011)
24. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)
25. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
26. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806–813 (2014)
27. Sünderhauf, N., Protzel, P.: BRIEF-Gist-Closing the loop by simple means. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1234–1241. IEEE (2011)
28. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4297–4304. IEEE (2015)
29. Tai, L., Liu, M., Deep-learning in mobile robotics-from perception to control systems: a survey on why and why not. arXiv preprint arXiv:1612.07139 (2016)
30. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
31. Weng, J., Zhang, Y., Hwang, W.-S.: Candid covariance-free incremental principal component analysis. IEEE Trans. Pattern Anal. Mach. Intell. **25**(8), 1034–1040 (2003)