Neurocomputing 390 (2020) 239–247

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Text-based indoor place recognition with deep neural network

Pei Li[a], Xinde Li[a,*], Hong Pan[a], Mohammad Omar Khyam[b,1], Md. Noor-A-Rahim[c]

[a] School of Automation, Key Laboratory of Measurement and Control of CSE, Ministry of Education, Southeast University, Nanjing, China
[b] Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA, USA
[c] School of Computer Science and IT, University College Cork, Cork, Ireland

ARTICLE INFO

Article history:
Received 29 November 2018
Revised 23 January 2019
Accepted 17 February 2019
Available online 26 October 2019

Keywords:
Indoor place recognition
CNN
LSTM
Natural language processing

ABSTRACT

Indoor place recognition is a challenging problem because of the hard representation to complicated intra-class variations and inter-class similarities.This paper presents a new indoor place recognition scheme using deep neural network. Traditional representations of indoor place almost utilize image feature to retain the spatial structure without considering the object's semantic characteristics. However, we argue that the attributes, state and relationships of objects are much more helpful in indoor place recognition. In particular, we improve the recognition framework by utilizing Place Descriptors (PDs) in text from to connect different types of place information with their categories. Meanwhile, we analyse the ability of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for classification in natural language, for which we use them to process the indoor place descriptions. In addition, we improve the robustness of the designed deep neural network by combining a number of effective strategies, *i.e.* L2-regularization, data normalization, and proper calibration of key parameters. Compared with existing state of the art, the proposed approach achieves well performance of 70.73%, 70.08% and 70.16% of accuracy, precision and recall on Visual Genome database respectively. Meanwhile, the accuracy becomes 98.6% after adding voting mechanics.

1. Introduction

Place recognition is one of the key issues in the semantic map research area. Its fundamental purpose is to enable service robots to perceive the environment via human understanding. In general, a specific place can be defined by the inside objects and a series of related tasks occurring within the objects. Therefore, place can be recognized through the positional relationships and attributes of objects or even people's state in the environment. To date, many approaches have been proposed from different aspects to address this issue [1]. However, most existing methods just focus on visual information itself without fully utilizing rich semantic contents in images. These image features only contain simple characteristics, *e.g.* texture, color, and geometric structure, so it is hard to accurately determine the type of place. On the other hand, the current research consider objects information only from single aspect, such as category, positional relationships,*etc.*, which still unable to simulate the process of human perception of the place. These above representation methods cannot describe a specific indoor place just like the definition. Meanwhile, there is currently no effective way to obtain the semantic characteristic of objects and it still face great difficulties. Therefore, it remains an open problem to extract semantic cue for place recognition, because semantic information is much more complex than visual information.

As for description model of object's attributes and relationships, witnessing the recent rapid development of deep learning application in image description and image captioning, some researchers tend to represent objects' information via a natural language model. For example, hierarchical recurrent network was presented in [2] to generate entire paragraphs for image description. Xu et al. [3] focused on the logical relationship between objects in the image, and generated more clear semantic words. Similar research [4,5] indicates that the attributes,status as well as relationships of objects in an image can be described by natural language model, which brings a new way to the object representation in place perception.

Natural language is another form of information representation that aligns well with human cognition process of things. It can ignore redundant information, and highlight the intrinsic attributes of the objects. Therefore, converting image information into text representation is beneficial for classification and inference. In this paper, we innovatively consider both the objects properties and positional relationships in the place, which has a more theoretical

* Corresponding author.
E-mail addresses: lipei_seu@seu.edu.cn (P. Li), xindeli@seu.edu.cn (X. Li).
[1] Current Affiliation: School of Engineering and Technology, Central Queensland University, Melbourne, Australia.

https://doi.org/10.1016/j.neucom.2019.02.065
0925-2312/

intuition than traditional methods that only considered image features. Particularly, our approach learns a priori text-based knowledge of object attributes, which is more helpful to judge the type of place.

Therefore, the contribution of this proposal is that we develop a deep learning-based approach which merges the image features of an indoor environment with the textual information converted from the image domain using LSTM-CNN. With such extra textual information, our new model does improve the recognition accuracy of indoor scene significantly.

This paper is organized as follows: Section 2 provides an overview of related work in indoor place recognition, especially the recent work using deep learning methods. In Section 3, we first analyse the basic scheme of place recognition, and then present an algorithm based on the LSTM-CNN for processing the objects information in text form. Section 4 introduces the implementation details of our algorithm, including data processing and model structure with training process. Section 5 reports experimental results of our methods, with a detailed comparison of other approaches. Finally, we conclude our work in Section 6.

2. Related work

Indoor place recognition is still a complicated problem due to the challenges of information integration and logical reasoning caused by high variability of indoor environment. It is generally considered that indoor place recognition mainly consists of three basic steps, *i.e.* image acquisition, information representation, and place classification. For information representation, different feature extraction methods have been proposed.

Recently, Place recognition has been extensively studied in the literature of computer vision. To address this problem, much effort has been devoted to empoly more visual information like texture, color, geometric information from infrared and/or laser,*etc.* For example, Mozoz et al. [6] applied Adaboost algorithm to data obtained from laser sensors and trained weak classifier sets to form a strong classifier. The boosted strong classifier enables the robot to perceive places like “room”, “corridor” and “doorway”. Swadzba et al. [7] studied 3D features capturing the spatial layout of indoor scenes on RGB-D image databases. In [8], Madokoro et al. presented an unsupervised scene classification method based on context of features which contained Visual Words (VWs) based on Scale-Invariant Feature Transform (SIFT) and Gist. In general, the above methods and related research [9,10,12,11] encoded visual and/or geometric information of the places into features which can be effectively used in place recognition. However, these methods only work for simply situations, but perform poor for scenes containing complex objects, especially for the open layout of the indoor place. This is mainly because the attributes and relationships of objects which play an important role in place classification are ignored in the feature extraction process. Therefore, some methods based on object information were proposed to tackle the above problem. In [13], Charalampous et al. fed the distribution of objects' location into a Naive Bayesian classifier to perform place categorization. Moreover, a Conditional Random Field (CRF) model was presented in [14] to jointly categorize objects and rooms from RGB-D images, exploiting object-object and object-room relations. In a similar way, Viswanathan et al. [15] used detection scores as well as learned object-place relations to perform place classification in images. The above research indicate that the accuracy of place recognition can be effectively enhanced by considering the properties of objects and their relationship. Therefore, motivated by these studies, our approach also involves the objects' properties and relationship in the pipeline of place recognition.

In addition to the above traditional methods, another important category methods of feature learning for indoor place information is based on deep learning. Deep learning models, especially deep neural networks, are capable of learning local feature representations as well as high-level abstractive features [16,17]. Since the indoor place category is related to the object properties and human's activities within the environment, deep learning methods are widely utilized in indoor place recognition. For example, a mid-level representation of convolutional features is proposed in [18,19]. In their papers, a deep convolutional neural network is developed for making use of the local cues, scene structure and object category relationships in images. Another example is the combination of GIST descriptor and discriminative deep belief network (DDBN) [20], in which DDBN extracts the non-linearly encoded information to form the global image feature. Essentially, since the deep neural networks are a series of non-linear mathematical operations, so these methods are powerful in automatically learning discriminative and high-level semantic feature representations that make the classification task much more easier in feature domain, rather than from the image information itself.

As human natural language is a symbolic abstractive information, many research focus on extraction of intrinsic feature from such symbolized information. Meanwhile, some work attempts to associate image information with natural language, because the latter are thought to be more closer to human cognition process. For example, recent work [21–23] *etc.* designed different deep learning structures and models to describe image contents into textual representation, some of which also contain image description of the object relationships and attributes [24].

3. The proposed approach

3.1. Overview of our method

As mentioned in Section 1, we attempt to utilize the natural language model to represent object information in images. Therefore, the place recognition are divided into three parts, as shown in Fig. 1. Firstly, for an image containing place with its categorical semantic label y, the object information such as objects status (denoted as set S) and attributes (denoted as set A), relationships (denoted as set R) is obtained by image description approach and/or by manual input through human machine interface. These information is called Place Descriptors (PDs) and can be expressed as $D \subset S \cup A \cup R$. Every element in the set D is in text-form. Since PDs provide an overall description of the object status, attribute and relationships with other objects, so they can be sufficiently used to predict the category of a place. Next as shown in Fig. 1(b), the above PDs,denoted as set D, are digitized into feature vector using a textual-digital transformation (denoted as function t). Thus, the digital information can be obtained and expressed as $D^{\star} = t(D)$. In particular, we utilize the Word2Vec method in our scheme. Finally, the set $(D^{\star}, y)$ containing all information of the place and its label are input into the developed LSTM-CNN-based model to perform the place classification.

3.2. Word2Vec transformation

Natural language is a kind of abstractive symbolic representation which is highly generalized and cannot be directly recognized by computers. For computer to effectively process natural language, the essential idea is to represent each language unit with a unique number and combine such numbers regularly to reflect rich language information. In this paper, the Word2Vec transformation is utilized to convert the English words into the numbers.

Word2Vec, also known as skip-gram model, was first proposed by Mikolov et al. [25,26]. It is an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. This word representation transfer one-hot

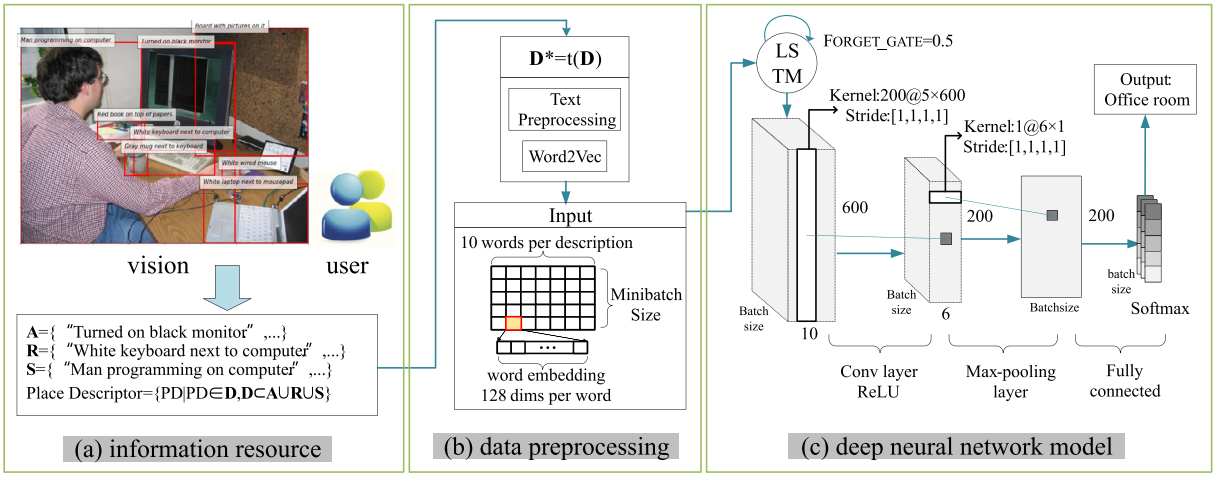

Fig. 1. Illustration of the architecture of our model. (a) place descriptors of objects in the specific place; (b) Digitization of descriptors; (c) LSTM-CNN-based classifier for semantic classification of indoor places.

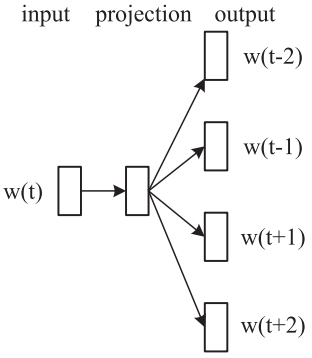

Fig. 2. The Skip-gram model architecture. The training object is to learn word vector representations that are good at predicting the nearby words.

encoded word vectors into distributed representations of words in a vector space, which helps to express the similarity and relationship between words.

The Skip-gram model aims to find word representations that are useful for predicting the surrounding words in a sentence. As shown in Fig. 2, given a sequence of training words $\omega_1, \omega_2, \omega_3, \ldots, \omega_T$, the objective of the Skip-game model is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(\omega_{t+j}|\omega_t), \tag{1}$$

where c is the size of the training context, which can be a function of the center word ω_t. The basic Skip-gram formulation defines $p(\omega_{t+j}|\omega_t)$ using the softmax function:

$$p(\omega_O|\omega_I)) = \frac{\exp\left({\upsilon'_{\omega_O}}^{\top}\upsilon_{\omega_I}\right)}{\sum_{\omega=1}^{W}\exp\left({\upsilon'_{\omega}}^{\top}\upsilon_{\omega_I}\right)}, \tag{2}$$

where υ_ω and υ'_ω are the "input" and "output" vector representations of ω, and W is the number of words in the vocabulary. When the value of the Eq. (1) satisfying the maximum requirements, the final value of the projection is the word vector representation of ω.

3.3. Place recognition with deep learning model

Based on the pipeline shown in Fig. 1, a LSTM-CNN model is established to solve the function t and perform place classification. For the classification model, because the PDs contain sufficient information and neural network models have demonstrated powerful capability in natural language processing (NLP) [27], CNN is adopted in our model to extract n-gram features at different positions in a sentence. Besides, the LSTM is also employed because it has great effect on handling sequences of any length sentences and capturing long-term dependencies. Finally, the hybrid deep neural network model, LSTM-CNN-based classifier, is applied for semantic classification.

4. Algorithm implementation

4.1. Data preprocessing

Since we focus on the semantic information of place, this model is assumed with the ability to recognize learn semantic representation. We utilize the dataset Visual Genome[1] [28] for generating the description corpus *i.e.* set D, which collect rich annotations of objects, attributes, and relationships within each image.

As shown in Fig. 1(a), we take the *bedroom* with its annotations as an example to illustrate the descriptors that we concerned. Here the objects in the image have been precisely marked and given descriptive texts. The annotations contain three different basic concepts of PDs, namely, the attributes of objects (e.g. "*Turned on black monitor*" in which 'monitor' is an object and 'turned on black' is its attributes, belonging to the set A); the positional relations between objects (e.g. "*Board with pictures on it*" in which 'Board' and 'pictures' are objects, 'on' is their positional relation, belonging to the set R) and the state of human (e.g. "*Man programming on computer*" in which 'programming' is human state, belonging to the set S). In addition, PDs may integrate the above basic concepts (e.g. "*White keyboard next to computer*" includes both object attributes and their positional relations, so it becomes $D \subset A \cup R$. Since human action and state play a crucial role in place recognition, we have to consider the state of human in the environment particularly. In most cases, PDs in dataset appear in the form of mixed concepts. Our model then learn from these PDs and generate a specific label. Therefore, if a new PD$\in D$ is processed by our model, the probable label y for an unknown place can be classified as the place in which contains the object learned before.

[1] https://visualgenome.org/

In order to ensure that these PDs can be converted to digital form by Word2Vec approach, we need to normalize them using data preprocessing algorithm, shown as below.

According to the Skip-gram model introduced in Section 3.2, the required parameter *wordsize* in Algorithm 1 represents the size

Algorithm 1 Data preprocessing.

Input: $[x_i]_k = \{w_n | w_n \subset \mathbf{T}, n = 1, \ldots, length(x_i)\}$ is the ith PD in the kth picture, w_n is the nth word in each PD, word vocabulary $\mathbf{T}$, *batchsize*, *maxlen*, *wordsize*.

Output: $[x_i^\star]_k$

1: **while** not end of k **do**
2: $m = \mathrm{mod}(length([x_i]_k), batchsize)$;
3: **if** $m > 0$ **then**
4: $[x_i]_k = [[x_s]_k, [x_i]_k], s = \mathrm{random}(1, i)$ and $length(s) = batchsize - m$;
5: **end if**
6: **for** $i \to 1; i \le length([x_i]_k)$ **do**
7: $[x_i]_k = \mathrm{normalize}([x_i]_k)$;
8: $[w_n^\star]_k = \mathrm{Word2Vec}([w_n]_k, wordsize, \mathbf{T}), [w_n]_k \in [x_i]_k$;
9: **if** $\max(n) < maxlen$ **then**
10: $[x_i^\star]_k = [[w_n^\star]_k, \mathrm{zeros}(wordsize, maxlen - n)]$;
11: **else**
12: $[x_i^\star]_k = [w_{1:maxlen}^\star]_k$;
13: **end if**
14: **end for**
15: **return** $[x_i^\star]_k$;
16: **end while**

of the training context in Eq. (2). In the 7th line of Algorithm 1, the normalization of PDs mainly includes the following steps:

(1) Remove punctuation, extra spaces, and special symbols that do not affect the original semantic cue of the PDs.
(2) Replace the numbers in the text with the corresponding words.
(3) Remove stop words without changing the original semantic cue.

In addition to normalization, other steps in Algorithm 1 are designed for the structure of the LSTM-CNN model. The *batchsize* is used to improve the generalization performance. The parameter *maxlen* represents the length of input sentence. Since *maxlen* is limited by the LSTM model, it must be fixed. As shown in Fig. 1(b), the mini-batch strategy is utilized in the proposed LSTM, which needs the input data to be a three-dimensional tensor. Each dimension represents the size of the word embedding, the length of the sentence, and the size of the batch respectively. If the PDs have different length, it will result in existence of empty elements in the input tensor which cannot be handled. Therefore, the sentence length should be consistent by adding zero tensor (zero tensor is a placeholder without representing any information) or splitting the sentence, which is to avoid existence of empty elements. Although PDs not necessarily have the same length, we normalize them in Algorithm 1 to satisfy the requirement of LSTM without information loss.

4.2. Model structure

The proposed semantic classifier of indoor places contains two types of neural networks i.e. LSTM and CNN as shown in Fig. 1(c).

LSTM [29] is a revised architecture of recurrent neural network (RNN) for handling sequences of any length and capturing long-term dependencies to avoid gradient explosion or vanishing in the standard RNN. In this paper, we apply a standard architecture of LSTM that consists of two time steps with one basic unit representing a piece of text synthetically.

At each time step, the output of the module is controlled by a set of gates as a function of the old hidden state h_{t-1} and the input at the current time step x_t. These gates are: the forget gate f_t, the input gate i_t, and the output gate o_t, respectively, and all of them decide how to update the current memory cell c_t and the current hidden state h_t. As for NLP, each step represents every position of word in the sentence. Because the dimension of the word vector is given as d, the dimension of the memory and other gates in LSTM share the same value. The LSTM transition functions are defined as follows:

$$\begin{cases} i_t = \sigma(W_i \cdot [h_{t-1}, x_t, c_{t-1}] + b_i), \\ f_t = \sigma(W_f \cdot [h_{t-1}, x_t, c_{t-1}] + b_f), \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t, c_t] + b_o), \\ q_t = \tanh(W_q \cdot [h_{t-1}, x_t] + b_q), \\ c_t = f_t \odot c_{t-1} + i_t \odot q_t, \\ h_t = o_t \odot \tanh(c_t). \end{cases} \tag{3}$$

where h_t is the expected result. σ is the logistic sigmoid function with an output in [0,1], $\tanh(\cdot)$ is the hyperbolic tangent function that has an output in $[-1, 1]$, and $\odot$ denotes the element-wise multiplication. Essentially, we can regard f_t as the function to control how much information from the old memory cell can be forgotten, i_t as the function to control how much new information can be stored in the current memory cell, and o_t decides what to output based on the memory cell c_t. Because LSTM can effectively integrate and memorize the features of sequence data, we input each word in the sentence to the LSTM with forget gate $o_t = 0.5$ step by step, and finally obtain a comprehensive expression of a sentence which is then fed into CNN.

CNN is a kind of multilayer feed-forward neural networks which consist of various combinations of convolutional layer, subsampling layer and fully connected layer. Thanks to its powerful capability of capturing local correlations of spatial or temporal structures, convolution becomes the central non-linear operation of CNN. Let k be the length of the filter, and vector $\mathbf{m} \in \mathbb{R}^{k \times d}$ denotes the filter for convolution operations. For each position j in the sentence, we have a window vector $\mathbf{w}_j$ with k consecutive word vector, given as:

$$\mathbf{w}_j = [\mathbf{x}_j, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_{j+k-1}], \tag{4}$$

Here, a filter $\mathbf{m}$ convolves with the window vectors at each position in a valid way to generate a feature map $\mathbf{c} \in \mathbb{R}^{L-k+1}$, and each element c_j of the feature map for window vector $\mathbf{w}_j$ is obtained as follows:

$$c_j = f(\mathbf{w}_j \odot \mathbf{m} + b), \tag{5}$$

where $\odot$ is element-wise multiplication, $b \in \mathbb{R}$ is a bias term and f is a nonlinear ReLU transformation function. Besides, our model uses multiple filters to generate different feature maps and a max-pooling is applied after the convolution to select the most important features. When we obtain the output from the max-pooling layer, a softmax function defined in Eq. (6) is added for classification.

$$\mathrm{P}(y = i|\mathbf{z}) = \mathrm{P}_i = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{C} e^{\mathbf{z}_j}}, \tag{6}$$

where P_i denotes the possibility of being classified to the ith class and C represents the number of classes. $\mathbf{z}_i$ is the output from the fully connected layer, *i.e.* $\mathbf{z} = \mathbf{W}\mathbf{c} + b$, which contains the whole information of a PD. Then the category corresponding to the maximum value of output in Eq. (6) indicates the category to which the object belongs. In this paper, we adopt a fundamental architecture to construct the classifier for semantic classification. As shown in Fig. 1(c), the architecture of our CNN includes three basic layers with their weights and bias. The first layer is a convolutional

layer with ReLU as its activation function, which receives the hidden state data (h_t) from LSTM; the second layer is a max-pooling layer that sub-sample the filtered data from convolutional layer; and the last layer is a fully-connected layer connecting the softmax function with its previous layer. After processed by the deep neural network, a description of the place from a image can be used to identify a specific place category.

In addition, the output of Eq. (6) cannot obtain the place category for the entire image. Therefore, a voting mechanism (VM) is added after acquiring category of every description. It counts the number of categories for all descriptions and classify the image to the place category that has the most votes.

4.3. Training approach

We train the entire model by minimizing the cross-entropy error, defined as follow:

$$E(\mathbf{x}^{(i)}, y^{(i)}) = \sum_{j=1}^{k} 1\{y^{(i)} = j\} log(\widetilde{y}_j^{(i)}), \tag{7}$$

In Eq. (7), the ith training data $\mathbf{x}^{(i)}$ and its true label $y^{(i)} \in \{1, 2, \ldots, k\}$ is given for learning, where k is the number of possible labels and have to be converted into one-hot vector in actual operation. The estimated probabilities $\widetilde{y}_j^{(i)} \in [0, 1]$ for each label $j \in \{1, 2, \ldots, k\}$ is the output of the softmax function. Besides, $1\{condition\}$ is an indicator such that $1\{condition\ is\ true\} = 1$, otherwise $1\{condition\ is\ false\} = 0$.

We employ the mini-batch gradient descent to learn the model parameters [30], more specifically, the parameter of mini-batch m in the 3rd line of Algorithm 2 corresponds to the *batchsize* in Algorithm 1. To avoid the overfitting problem, the learning rate is reduced gradually when exceeding a given epoch. Besides, we use L2 regularization approach to ensure parameters have reasonable values. The basic process of our training algorithm is shown in Algorithm 2. During training process, every weight and bias parameter ($\boldsymbol{\theta}$) in the neural network are updated until reaching the maximum iteration step or the error is below a threshold.

Algorithm 2 Mini-batch gradient descent with L2 regularization.

Input: Learning rate lr_k, learning decay rate d, max decay epoch τ, regularization coefficient λ, initial θ;
Output: Updated parameter θ;

```
k ← 1;
while stopping criterion not met do
    Sampling a mini batch of m examples from the training set {x^(1), ..., x^(m)} with corresponding targets y^(i);
    Compute the mean square error of cross-entropy with L2 regularization: L((x^(i); θ), y^(i)) = (1/m) Σ ||E(x^(i), y^(i))||^2 + (λ/2m) Σ ||θ||^2;
    Compute gradient estimate: ĝ ← ∇_θ L((x^(i); θ), y^(i));
    Apply update: θ ← θ − lr_k × ĝ;
    if k > τ then
        lr ← lr × d^(k−τ);
    end if
    k = k + 1;
end while
```

5. Experiment

5.1. Parameter settings

In our experiment, we choose five indoor categories including *kitchen, bedroom, living room, bathroom* and *office*, and each of them contains 50 images from Visual Genome. These contents of annotations are the prior knowledge which our model needs to learn. Table 1 shows the statistics of the dataset and some examples for each category are illustrated in Fig. 3.

Table 1
Statistics of the Visual Genome dataset.

Category	Number of pictures	Number of PDs	One-hot encoder
Bathroom	50	3149	[1,0,0,0,0]
Bedroom	50	2597	[0,1,0,0,0]
Kitchen	50	2929	[0,0,1,0,0]
Living room	50	3081	[0,0,0,1,0]
Office	50	3978	[0,0,0,0,1]
Total	250	15,734	-

According to the instructions of Word2Vec[2], there are four key parameters in the function, namely, dimensionality of the word vectors (size = 128), maximum distance between the current and predicted word within a sentence (window = 5), ignores all words with total frequency lower than this (mincount = 1), number of iterations (epochs) over the corpus (itera = 50). We have obtained the above preferable values through many experiments.

After processed by Algorithm 1, every different word in the dataset can be converted into a unique 128-dimensional vector which is applied to represent the normalized digital PDs. Finally, the parameter *maxlen* is set as 10, which means that there are 10-word length in every PDs. Furthermore, in order to facilitate classifier operation, each category is converted into an unique string of numbers to represent the label by one-hot encoding method.

When using LSTM for natural language processing, the parameter t represents the position of words in a text sequence. In Eq. (3), $\mathbf{x}_t \in \mathbb{R}^d$, where $d = 128$ means the dimension of word vector mentioned above. Besides, we set the value of forget gate to 0.5, while taking the architecture into account. This indicates that half of the old information from output data in previous step is forgotten. The value of input and output gate are set as 1. Usually, high dimensional word vectors can encode rich information, therefore we let the number of hidden state neuron (h_t) be 600 and each gate (o_t, i_t, f_t) as well as state cell(c_t, q_t) set as 200 to ensure that text features can extract as much semantic information as possible. So far all fundamental parameters about the architecture of LSTM have been set. After a fixed step size (*maxlen*) processed by LSTM, we can obtain a 200-dimensional vector to represent the information of a phrase.

For parameters in CNN, more specifically, we utilize the mini-batch technique to improve its generalization ability and set the batch size to 32, so that it transforms the shape of the data into $10 \times 600 \times 32$. Then the convolutional layers convolve the input data with 200 kernels of size 5×600 with a stride of 1 step for every batch. The Max-pooling layer has one kernel of size 6×1 for every batch. Then, the 200−dimensional data are fully connected to the last 5 neurons. Finally, the output of the last fully connected layer is fed to a 5-way softmax layer, which produces a distribution over 5 class labels.

5.2. Experiment results

In this section, we verify the performance of our model. Firstly, the dataset mentioned in Section 5.1 is randomly divided into two parts, with 70% as training set and the remaining 30% as test set. Based on Algorithm 2, we tested our model many times and choose the model parameters as follow: learning rate $lr_k = 0.15$, learning decay rate $d = 0.95$ and maximum decay epoch $\tau = 0.15$.

[2] https://radimrehurek.com/gensim/models/word2vec.html

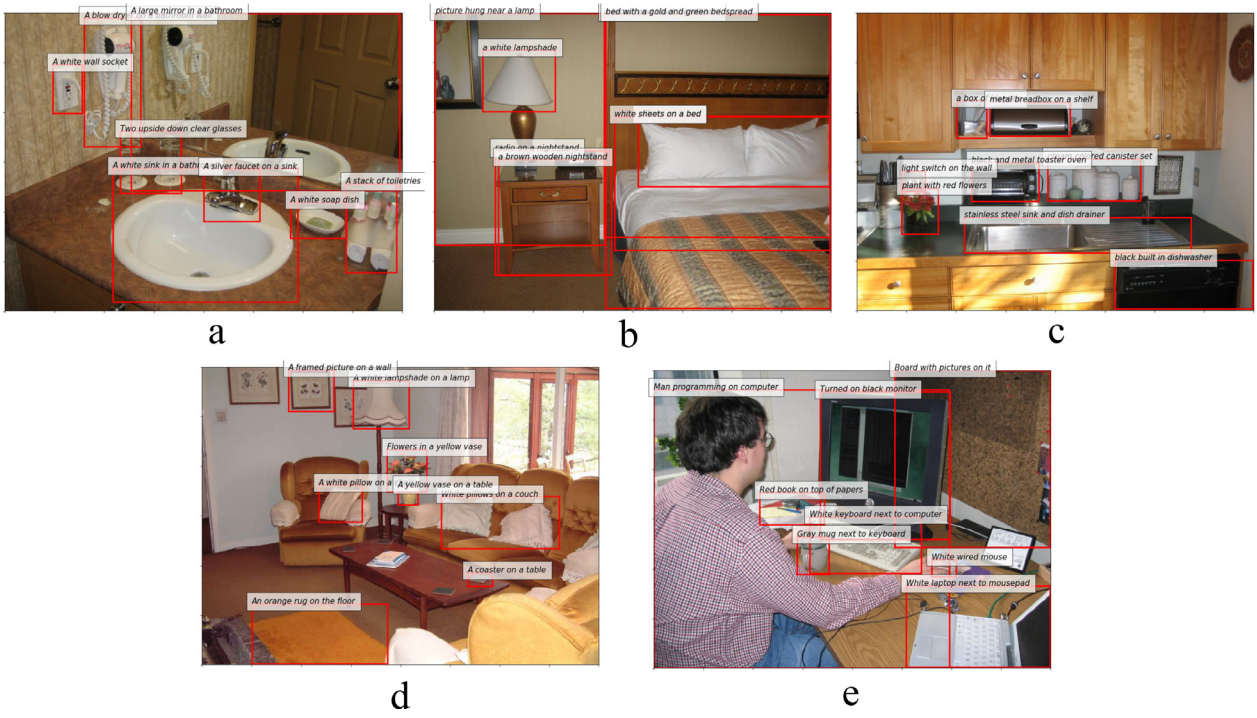

Fig. 3. Examples of indoor places with their PDs: (a)Bathroom, (b)Bedroom, (c)Kitchen, (d)Living room, (e)Office.

Table 2
Confusion matrix of the indoor place recognition for every PDs. (Note: The labels corresponding to: Bathroom(i), Bedroom(ii), Kitchen(iii), Living room(iv), and Office(v)).

Confusion Matrix		Predicted category				
		i	ii	iii	iv	v
Actual category	i	900	48	179	55	34
	ii	60	653	42	219	50
	iii	123	28	766	69	38
	iv	45	170	123	744	166
	v	38	43	66	109	1056

Table 3
Precision, recall, F1 score per category and accuracy.

Category	Precision(%)	Recall(%)	F1 score(%)	Accuracy(%)
Bathroom	77.19	74.01	75.57	-
Bedroom	69.32	63.77	66.43	-
Kitchen	65.14	74.80	69.64	-
Living room	62.21	59.62	60.88	-
Office	78.57	80.49	79.52	-
Average	70.48	70.54	70.41	70.72
std(standard deviation)	(±6.46)	(±7.67)	(±6.58)	

Table 4
Confusion matrix of the indoor place recognition for every picture. (Note: The labels corresponding to: Bathroom(i), Bedroom(ii), Kitchen(iii), Living room(iv), and Office(v)).

Confusion Matrix		Predict category				
		i	ii	iii	iv	v
Actual category	i	14	0	1	0	0
	ii	0	14	0	1	0
	iii	0	0	14	1	0
	iv	0	0	0	15	0
	v	0	0	0	0	15

Table 5
Accuracy and training time in different proportion of test set.

Proportion	Accuracy (%) per description	Accuracy (%) per picture	Training time (sec)
0.1	69.8	96.0	10,244
0.2	70.1	98.0	7246
0.3	70.7	96.0	6581
0.4	70.6	96.0	5412
0.5	71.4	97.6	4783
0.6	69.7	98.0	3653
0.7	67.7	97.1	2765
0.8	67.7	95.5	1877
0.9	64.5	95.5	1024

The programming environment is TensorFlow v1.4 and Python v3.6 installed on a Intel i5 CPU with 8 GB memory platform.

Table 2 illustrates the results in terms of the confusion matrix, and we report precision, recall and F1 score [31] for each category as well as overall average accuracy in Table 3. Experimental results show that our algorithm has a certain recognition accuracy for a single PDs. Based on this observation, we consider processing PDs from the same image jointly. Once the results from a single PD's category for test samples are available, a voting mechanism (VM) is added to determine the final category of the place, which is also satisfied with the basic logic of place recognition tasks. Table 4 lists the overall recognition result in format of confusion matrix. As shown in Table 4, our method finally achieves an average accuracy of 96%.

5.3. Discussion

In addition to evaluating the effectiveness of our algorithm, we also verified the generalization and uncertainty performance of our algorithm. Table 5 shows the accuracy in different parameter settings, where the first column denotes the proportion of the test set in the total number of samples. In this experiment, we gradually reduced the number of training samples. As we can see, although

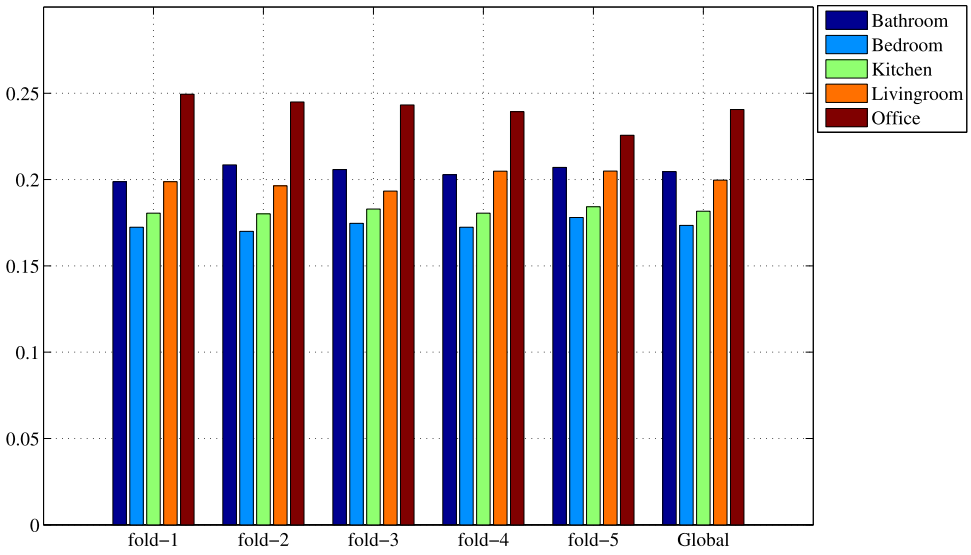

Fig. 4. Distribution of the dataset categories in the 5-fold cross-validation and global distribution of all training set.

Table 6
Evaluation results for per PDs in the first four rows and per picture in the last row. Each column shows the average and standard deviation for $k = 5$ executions of the k-fold cross-validation.

	fold-1	fold-2	fold-3	fold-4	fold-5	Average
Precision	72.50 ± 6.36	71.16 ± 5.39	66.87 ± 6.92	72.26 ± 8.01	67.60 ± 12.95	70.08 ± 2.38
Recall	72.61 ± 7.31	71.28 ± 9.66	67.00 ± 9.24	72.02 ± 8.11	67.87 ± 9.90	70.16 ± 2.28
F1 score	72.36 ± 5.55	71.14 ± 7.37	66.90 ± 8.00	72.10 ± 7.86	67.63 ± 11.16	70.03 ± 2.30
Accuracy	72.59	71.50	67.21	73.07	69.26	70.73 ± 2.19
VM accuracy	98.00	100.00	92.00	96.00	98.00	96.8 ± 2.71

the count of correctly classified training samples are decreasing, the overall recognition accuracy rate of test samples still maintains at a high level. It indicates that our algorithm has a certain fault tolerance. We argue that this is mainly due to the generalization performance of the neural network, which automatically extracts the most discriminative features of the PDs. For example, the indoor place of bathroom often contains PDs such as "*the sink is white*", these PDs play a major role in the place recognition. As the training samples become less, the indoor place information learned by the neural network also reduce at the same time, so the classification accuracy rate for every PDs slightly decrease. On the other hand, because some of the discriminative features are "remembered" by the neural network, which can still guarantee the correct classification of images with the help of the voting mechanism.

Besides, the proposed method is evaluated using a 5-fold cross-validation procedure. To perform a fair comparison among dataset, the cross-validation folds in all evaluations have 10 test samples and 40 training samples for each category. As shown in Fig. 4, place categories are not uniformly distributed across all folds. Mean precision, accuracy, F1 score, accuracy and accuracy with voting mechanics as well as the standard deviation are reported in Table 6 for each fold. High standard deviation are obtained because of the extreme variation in the categories distribution among folds (shown in Fig. 4). As mentioned before, the generalization performance can be acceptable in cross-validation method. Therefore, our method using PDs in text-form is robust to a certain extent.

Meanwhile, we also analyse a representative example that is recognized incorrectly. Fig. 5 shows two indoor place images together with their PDs. Taking Fig. 5(a) as a example, this image is manually marked as bathroom but our method predicted it computes the predicted label as kitchen. The details of experimental results are reported in the Table 7. We find that if an indoor place equipped with complicated objects or has shared attributes (*e.g.*

Table 7
An example (Figure 5(a)) of experimental results. The actual label is bathroom, but the predicted label is kitchen. (Parts of PDs' predicted results are omitted because of space limitation).

PDs	Predicted Results for each PDs	Total
a white porcelain sink basin a silver metal drain stopper a white plastic lid a handle on a faucet The counter is white *etc.*	Bathroom	23
a table in the room a chair in the room vertical mini blinds in window The chair is near the window wooden chair next to a table *etc.*	Bedroom	10
a black coffee machine glass jars on the counter the faucet on the sink the sink on the counter white mugs on the counter *etc.*	Kitchen	24
burgandy cushion on chair Glasses are stacked on the counter A wooden chair near a table A table near a window wooden table near a window *etc.*	Livingroom	10
a white coffee mug The coffee mug is white fan is white top coffee mug of a stack	Office	4

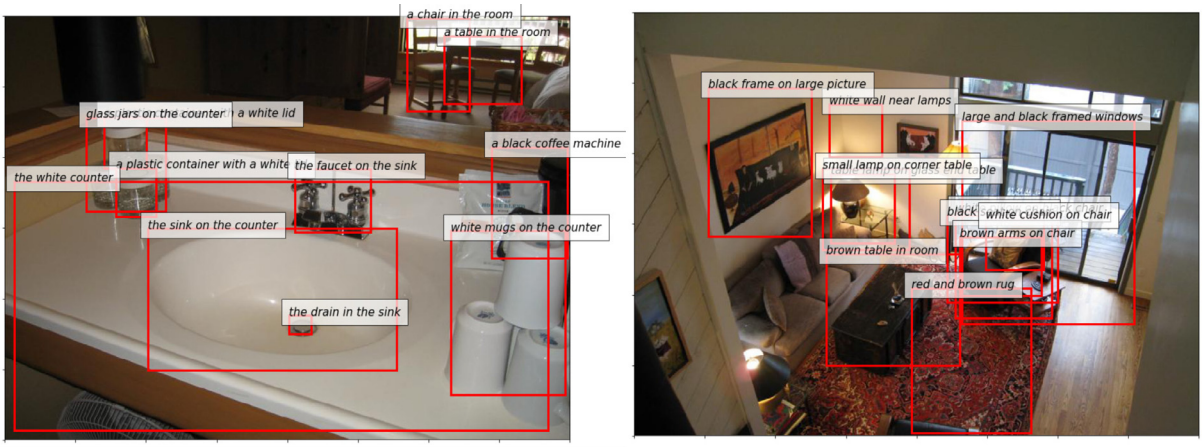

Fig. 5. Two examples of incorrect recognition results. (a) predicted label: Kitchen, actual label: Bathroom. (b) predicted label: Office, actual label: Living room. (Parts of PDs are omitted because of space limitation).

Table 8
Comparison of different feature representation methods (ϕ_6 and ϕ_7 represent the results of our method without and with the VM, respectively).

Methods	ϕ_1[10]	ϕ_2[32]	ϕ_3[33]	ϕ_4 [20]	ϕ_5[34]	ϕ_6	ϕ_7
Accuracy	85.5%	84%	73.36%	69.43%	81.2%	70.72%	96.8%

"*the sink is white*" probably appears in the kitchen and the bathroom), it will make our method difficult to predict the correct category.

We compare the proposed PD descriptor with the other different information representation methods, Table 8 shows the comparison results. Taking visual clues as an example, the highest recognition accuracy is 85.5% in Ranganathan et al.'s research [10], and average accuracy is 46.85% for six indoor places' category. Besides, Swadzba et al. [32] combined different kinds of image features to determine the place category and the classification rates is 84% on average. Romero et al. [33] added the depth information based on the 3D spatial pyramid to generate indoor scene descriptors from point clouds. Their method achieved 73.36% in classification accuracy. Furthermore, the methods [20] and [34] presented two kinds of deep learning model called CNN and DDBN, respectively to generate image features. As seen in Table 8, our method leads other methods by a large margin in terms of recognition accuracy.

6. Conclusion

This paper presents a strategy for text-based indoor place recognition which applies LSTM-CNN network structure. Classification accuracy and other quantitative metrics of our method are evaluated experimental results indicate that our approach is effective in the place recognition problem. For the future work, We would focus on exploring the semantic connection of the objects and human state holistically in one place. Besides, the generalization performance of the proposed model needs further research and improvement. We believe that the indoor scene recognition will benefit from this indirect solution.

Declaration of Competing Interest

None

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61573097, 61671151 and 91748106), in part by Key Laboratory of Integrated Automation of Process Industry (PAL-N201704), the Natural Science Foundation of Jiangsu Province (BK20181265), the Fundamental Research Funds for the Central Universities (3208008401), the Qing Lan Project and Six Major Top-talent Plan, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J.J. Leonard, D. Cox, P. Corke, M.J. Milford, Visual place recognition: a survey, IEEE Trans. Robot. 32 (1) (2016) 1–19, doi:10.1109/TRO.2015.2496823.
[2] J. Krause, J. Johnson, R. Krishna, L. Fei-Fei, A hierarchical approach for generating descriptive image paragraphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3337–3345.
[3] D. Xu, Y. Zhu, C.B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2, 2017.
[4] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 852–869.
[5] L. Shen, S. Yeung, J. Hoffman, G. Mori, L. Fei-Fei, Scaling human-object interaction recognition through zero-shot learning, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1568–1576.
[6] O.M. Mozos, C. Stachniss, W. Burgard, Supervised learning of places from range data using adaboost, in: Proceedings of the IEEE International Conference on Robotics and Automation,ICRA, IEEE, 2005, pp. 1730–1735, doi:10.1109/ROBOT.2005.1570363.
[7] A. Swadzba, S. Wachsmuth, A detailed analysis of a new 3D spatial feature vector for indoor scene classification, Robot. Autonom. Syst. 62 (5) (2014) 646–662, doi:10.1016/j.robot.2012.10.006.
[8] H. Madokoro, Y. Utsumi, K. Sato, Scene classification using unsupervised neural networks for mobile robot vision, in: Proceedings of the SICE Annual Conference (SICE), 2012, 2012, pp. 1568–1573.
[9] L. Shi, S. Kodagoda, G. Dissanayake, Multi-class classification for semantic labeling of places, in: Proceedings of the 11th International Conference on Control Automation Robotics & Vision (ICARCV), 2010, pp. 2307–2312, doi:10.1109/ICARCV.2010.5707856.
[10] A. Ranganathan, Pliss: labeling places using online changepoint detection, Autonom. Rob. 32 (4) (2012) 351–368, doi:10.1007/s10514-012-9273-4.
[11] C. Romero-Gonzlez, J. Martnez-Gmez, I. Garca-Varea, L. Rodrguez-Ruiz, On robot indoor scene classification based on descriptor quality and efficiency, Expert Syst. Appl. 79 (C) (2017) 181–193, doi:10.1016/j.eswa.2017.02.040.
[12] Y. Zheng, J. Pu, H. Wang, H. Ye, Indoor scene classification by incorporating predicted depth descriptor, in: B. Zeng, Q. Huang, A. El Saddik, H. Li, S. Jiang, X. Fan (Eds.), Advances in Multimedia Information Processing – PCM 2017, Springer International Publishing, 2018, pp. 13–23, doi:10.1007/978-3-319-77383-4_2.

[13] K. Charalampous, I. Kostavelis, F.-E. Chantzakou, E.-S. Volanis, C. Emmanouilidis, P. Tsalides, A. Gasteratos, Place categorization through object classification, in: Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), 2014, pp. 320–324, doi:10.1109/IST.2014.6958497.
[14] J.-R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez, Joint categorization of objects and rooms for mobile robots, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 2523–2528, doi:10.1109/IROS.2015.7353720.
[15] P. Viswanathan, T. Southey, J.J. Little, A. Mackworth, Automated place classification using object detection, in: Proceedings of the Canadian Conference on Computer and Robot Vision (CRV), 2010, pp. 324–330, doi:10.1109/CRV.2010.49.
[16] A. Yildiz, N. Öztürk, N. Kaya, F. Öztürk, Integrated optimal topology design and shape optimization using neural networks, Struct. Multidiscipl. Optim. 25 (4) (2003) 251–260, doi:10.1007/s00158-003-0300-0.
[17] N. Öztürk, A.R. Yıldız, N. Kaya, F. Öztürk, Neuro-genetic design optimization framework to support the integrated robust design optimization process in ce, Concurr. Eng. 14 (1) (2006) 5–16.
[18] S.H. Khan, M. Hayat, M. Bennamoun, R. Togneri, F.A. Sohel, A discriminative representation of convolutional features for indoor scene recognition, IEEE Trans. Image Process. 25 (7) (2016) 3372–3383, doi:10.1109/TIP.2016.2567076.
[19] M. Hayat, S.H. Khan, M. Bennamoun, S. An, A spatial layout and scale invariant feature representation for indoor scene classification, IEEE Trans. Image Process. 25 (10) (2016) 4829–4841, doi:10.1109/TIP.2016.2599292.
[20] N. Zrira, H.A. Khan, E.H. Bouyakhf, Discriminative deep belief network for indoor environment classification using global visual features, Cognit. Comput. 10 (3) (2018) 437–453, doi:10.1007/s12559-017-9534-9.
[21] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3128–3137, doi:10.1109/CVPR.2015.7298932.
[22] K. Fu, J. Li, J. Jin, C. Zhang, Image-text surgery: efficient concept learning in image captioning by generating pseudopairs, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5910–5921, doi:10.1109/TNNLS.2018.2813306.
[23] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651–4659, doi:10.1109/CVPR.2016.503.
[24] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[25] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Comput. Sci. (2013).
[27] C. Zhou, C. Sun, Z. Liu, F.C.M. Lau, A c-lstm neural network for text classification, Comput. Sci. 1 (4) (2015) 39–44.
[28] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.
[29] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, doi:10.1162/neco.1997.9.8.1735.
[30] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, 1, MIT press Cambridge, 2016.
[31] D. Powers, Evaluation: From precision, recall and fmeasure to roc, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2007) 37–63.
[32] A. Swadzba, S. Wachsmuth, Indoor scene classification using combined 3D and gist features, in: Proceedings of the Asian Conference on Computer Vision, 2010, pp. 201–215.
[33] C. Romero-González, J. Martínez-Gómez, I. García-Varea, L. Rodríguez-Ruiz, 3D spatial pyramid: descriptors generation from point clouds for indoor scene classification, Mach. Vis. Appl. 27 (2) (2016) 263–273, doi:10.1007/s00138-015-0744-4.
[34] M. Hayat, S.H. Khan, M. Bennamoun, S. An, A spatial layout and scale invariant feature representation for indoor scene classification, IEEE Trans. Image Process. 25 (10) (2015) 4829–4841, doi:10.1109/TIP.2016.2599292.

Pei Li is currently studying the Ph.D. and Master degree of Control Science and Engineering at the School of Automation, Southeast University. He received his Bachelor degree with a major in Automation from the Tongji University. His main research include place perception, deep learning, natural language processing.

Xinde Li received his Ph.D. from the Department of Control, Huazhong University of Science and Technology in June 2007. In December of the same year, he worked in the School of Automation, Southeast University. From January 2012 to January 2013, he visited Georgia Polytechnic University as a national visiting scholar for one year. From January 2016 to the end of August 2016, he worked as a research fellow in the Department of ECE, National University of Singapore. His main research interests include intelligent robots, machine vision perception, machine learning, human-computer interaction, intelligent information fusion and artificial intelligence.

Hong Pan is associate researcher at the School of Automation, Southeast University. In 2004, he graduated from Southeast University with his Ph.D. in pattern recognition and intelligent systems. From September 2004 to August 2006, he worked as a research associate at the Multimedia Signal Processing Center of the Hong Kong Polytechnic University, where he worked on image sparse transform and coding. His research interests include machine learning, deep learning, computer vision, medical image processing and analysis, multimedia signal processing (image/video codec, retrieval and analysis).

Mohammad Omar Khyam received the B.Sc. degree in electronics and telecommunication engineering from the Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, in 2010, and the Ph.D. degree from the University of New South Wales, Australia, in 2015. He is currently a Lecturer with the Central Queensland University, Australia. His research interests include signal processing and wireless communication.

Md. Noor-A-Rahim received the Ph.D. degree from Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia, in 2015. He was a Postdoctoral Research Fellow with the Centre for Infocomm Technology, Nanyang Technological University, Singapore. He is currently a Senior Postdoctoral Researcher (MSCA Fellow) with the School of Computer Science and IT, University College Cork, Cork, Ireland. His research interests include control over wireless networks, intelligent transportation systems, information theory, signal processing, and DNA-based data storage. He was the recipient of the Michael Miller Medal from the Institute for Telecommunications Research, University of South Australia, for the most Outstanding Ph.D. Thesis in 2015.