Hindawi
Mobile Information Systems
Volume 2020, Article ID 4835198, 14 pages
https://doi.org/10.1155/2020/4835198

Research Article

Research on Indoor Scene Classification Mechanism Based on Multiple Descriptors Fusion

Ping Ji, Danyang Qin, Pan Feng, Tingting Lan, and Guanyu Sun

Key Lab of Electronic and Communication Engineering, Heilongjiang University, Harbin, China

Correspondence should be addressed to Danyang Qin; qindanyang@hlju.edu.cn

Received 25 August 2019; Accepted 22 January 2020; Published 16 March 2020

Guest Editor: Malik Jahan Khan

This study aims at the great limitations caused by the non-ROI (region of interest) information interference in traditional scene classification algorithms, including the changes of multiscale or various visual angles and the high similarity between classes and other factors. An effective indoor scene classification mechanism based on multiple descriptors fusion is proposed, which introduces the depth images to improve descriptor efficiency. The greedy descriptor filter algorithm (GDFA) is proposed to obtain valuable descriptors, and the multiple descriptor combination method is also given to further improve descriptor performance. Performance analysis and simulation results show that multiple descriptors fusion not only can achieve higher classification accuracy than principal components analysis (PCA) in the condition with medium and large size of descriptors but also can improve the classification accuracy than the other existing algorithms effectively.

1. Introduction

With the rapid development of the Internet and the increasing demand for applications based on location awareness, location-based services are getting extensive attention. Most people cannot live without the location service and the navigation system based on GPS (Global Position System) in their daily life. Obviously, outdoor localization technology has been relatively mature, and many mobile devices also refer to outdoor location technology [1, 2, 3, 4]. Due to the particularity of indoor environment, the GPS signal cannot directly meet the requirements of indoor localization service. At present, there are many indoor localization methods [4–6], mainly including WiFi, RFID, Bluetooth, Ultrawide band, and so on. Nowadays, the visual indoor localization system [7–9] is attracting more and more attentions of the researchers all over the world due to the advantages of low deployment cost, strong autonomy, and high localization accuracy.

A large visual database, namely, Visual Map, has occasionally been established at offline stage to achieve accurate indoor visual localization. Visual Map may contain a large number of images or image features of different scenes and corresponding location information, which is the foundation of visual indoor localization. When the user performs a location query online, the image will be retrieved in the Visual Map. Traditional image retrieval algorithms rely on pixel point matching [10, 11], which can only give the results of image matching but does not contain the visual image location information. In addition, existing image retrieval algorithms often carry out global traversal search, which leads to excessive time overhead and is not conducive to real-time localization of mobile users. Therefore, an effective indoor scene classification mechanism is proposed in this paper based on multiple descriptors fusion. The images in Visual Map will be classified according to the scenes, so as to reduce the time overhead of visual images retrieval at online stage and improve the efficiency and accuracy of indoor scene classification. In this paper, both the visual information and the depth information of an image are fused. The visual image mainly contains color information, and each point on the depth image corresponds to the visual image and contains position information. Both types of images are captured by Microsoft Kinect 2.0.

In the indoor scene classification mechanism, the initial descriptor set containing two kinds of image descriptors will

be generated by the existing spatial pyramid model (SPM) [12, 13]. Then, the greedy descriptor filter algorithm (GDFA) will be proposed to find out the valuable descriptors. Multiple fusion descriptors will be generated by homologous and nonhomologous combination to further enhance the effectiveness of descriptors. Finally, support vector machine (SVM) will be adopted for classification. The overall framework of the indoor scene classification mechanism is shown in Figure 1.

The remaining of the paper is arranged as follows: Section 2 reviews the research progress of scene classification techniques and their applications in indoor scenes. Section 3 describes the generation of the initial descriptor set and the descriptor filtering in detail. Section 4 introduces the experimental database of this paper and shows descriptor evaluation results. In Section 5, two combinations of homologous and nonhomologous will be realized and the combination results will be evaluated. Section 6 concludes the article.

2. Motivation

At the Scene Understanding Symposium held at MIT in 2006, an important point was clearly stated for the first time, namely, scene classification is a new promising research direction for image understanding. Although existing classification methods claim to be able to solve any scene classification problems [14, 15], the experimental outcome shows that only the outdoor scene classification can be effectively solved by these methods, while the indoor scene classification problems may still be a challenging task. In addition, [16] shows that the classification accuracy of the indoor scene is far lower than that of the outdoor scene adopting the same feature extraction and classification recognition methods. Therefore, it is important to improve the classification accuracy of the indoor scene.

In early studies, low-level features of images were usually extracted to classify scenes, such as color, texture, and shape [17–19]. However, these methods based on low-level features have not been a hot topic in the field of scene classification due to its unsatisfactory classification effect. In order to overcome such problems, the methods based on middle-level features of image are proposed. The global feature Gist is adopted and improved in [20]. The good identification ability of scale invariant feature transform (SIFT) makes it always be adopted as the local features with the highest priority in many scene recognition algorithms [21]. Shi et al. [22] proposed an indoor scene classification algorithm based on the enhancement of visual sensitive area information. And local features and global features are integrated by the visual sensitive area information.

With the rise of Kinect, the scene classification algorithm based on depth information [24, 25] has received more and more attention. The histogram of oriented gradient (HOG) algorithm [26] is adopted to classify depth images and visual images, respectively [28]. SIFT is adopted to extract features of depth images and color images, and SPM coding is adopted to classify images after feature fusion [29]. SIFT of visual images and speeded up robust features (SURF) [27] of depth images are fused to classify images [30]. Five deep core feature extraction algorithms are designed in [31] to extract the size, edge, and shape information of visual images, respectively, and the extracted information is fused for classification.

As research continues, the model based on the convolutional neural network (CNN) [16, 23] has attracted the researchers. However, massive training sets are required in CNN, which may result in relatively long training time. In addition, CNN usually has high computing requirement on the platform, so it is difficult to realize indoor scene classification on the platform with limited computing resource.

3. Multiple Image Descriptor Generation and Filtering

Inspired by [28–31], visual information and depth information will be fused in this paper. The higher accuracy indoor scene classification effect will be achieved by the spatial 3D information contained in the depth image, which is insensitive to light and reflects the position relationship between objects. Features of the original images will be extracted by D-SIFT (Dense SIFT) [32], and similar features will be clustered to form BoW (Bag-of-Words) [33–35] by K-means [36, 37]. Based on BoW, the initial descriptors set including visual image descriptors and depth image descriptors will be generated with the construction of SPM. It is true that the number of initial descriptors is large and the quality is uneven. In addition, combining directly with unfiltered initial descriptors will lead to an explosion of the combined results. Therefore, a simple and effective descriptor filtering algorithm ought to be proposed to obtain those valuable descriptors.

3.1. Initial Descriptors Generation. The descriptor generated expression could be derived from the following procedure. Let I be any input image and x be a descriptor generated by the image. $\mathbb{L}$ is a set of predefined class tags, and l is one of them. The function of generating descriptor x from image I can be expressed as $g(I) = x$, and the probability of successfully matching descriptor x to class tag l is $Pl \mid x$. Therefore, the expression of the most appropriate class tag $\widetilde{l}$ will be

$$\widetilde{l} = \arg\max_{l \in \mathbb{L}} \mathrm{P}(l \mid g(I)). \tag{1}$$

The key to the research will be turning the initial descriptors into valuable descriptors with high classification accuracy. In order to find such descriptors, equation (1) will be further optimized. On the premise of the best descriptor filtering and combination methods, a correct class label assigned to input image I will be $\widehat{l}$ ($\widetilde{l} \neq \widehat{l}$) and $\mathcal{X}$ is adopted to express a set of multiple image descriptors. Then, the optimized descriptor generation expression will be

$$\widetilde{g}(I) = \arg\max_{g(I) \in \mathcal{X}} P(\widehat{l} \mid g(I)). \tag{2}$$

According to equation (2), the initial descriptors generated by the input image can only get the desired

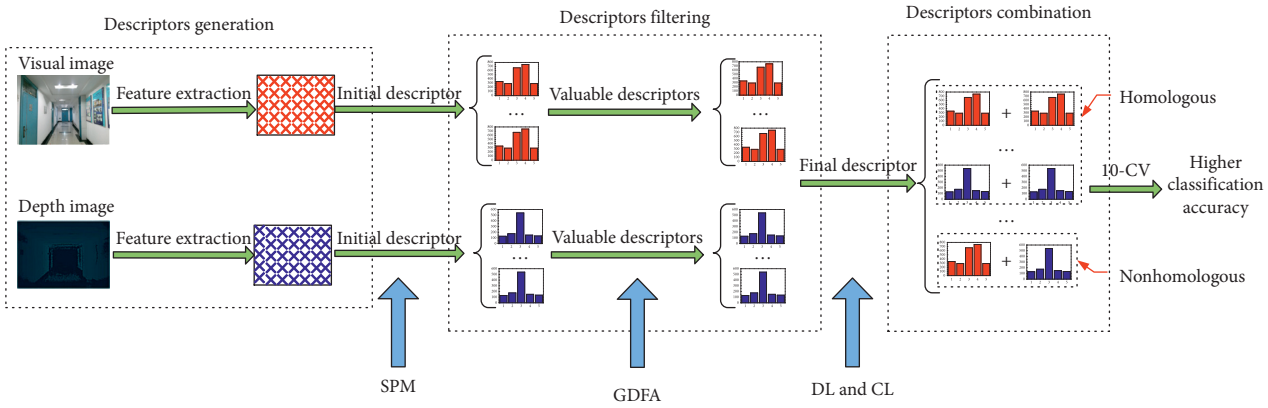


Figure 1: The indoor scene classification mechanism.

classification effect through filtering and combination. Initial descriptors are large in number and poor in quality, while descriptor filtering can discard worthless descriptors and descriptor combination can improve the effectiveness of descriptors. The descriptor generation process based on SPM will be described as follows.

3.2. Spatial Pyramid Model. In recent years, the BoW model has been widely adopted in computer vision. It takes the image features as visual words and classifies images by counting the number of visual words in each image. However, the traditional BoW lacks the spatial position information [29]. In this research, SPM will be established to cut the image into scale cells, then the number of visual words will be counted in each cell and the histograms can be drawn. Finally, histogram features at all scales will be linked together to form an eigenvector. We assume that a part of visual words has been selected as basic features. The steps of descriptor generation based on SPM are described in detail as follows:

(i) Extracting the D-SIFT feature.

(ii) Mapping each feature point to the corresponding visual word.

(iii) Cutting the image and constructing spatial pyramid hierarchy (three cutting methods, such as vertical cutting method, horizontal cutting method, and grid cutting method, are adopted in this paper, as shown in Figures 2(a)–2(c), respectively).

(iv) Counting the number of visual words in each cell and plotting histograms for each cell.

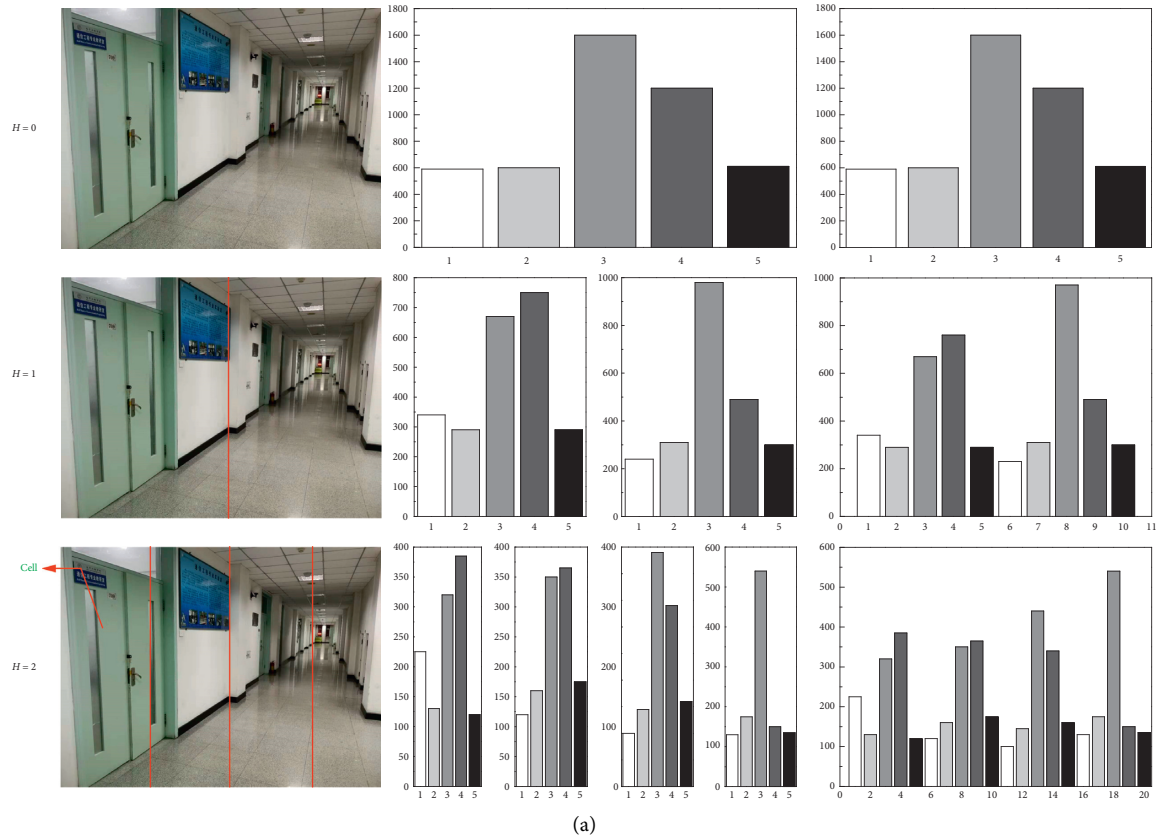
(v) Connecting all histograms to form a feature vector as the image descriptor.

The SPM-based descriptor generation process is shown in Figures 2(a)–2(c), and each cutting type will be divided into three columns for clear explanation. As shown in Figure 2(a), the first column shows the cutting type of the initial image, the second column represents the statistical results of visual words for each cell, and the initial descriptors formed by connecting the second column histograms are shown in the third column. The image contains 5 visual words; three pyramid hierarchies; and vertical, horizontal, and grid, the three cutting methods. The descriptors generation based on SPM mainly depends on three important parameters: BOW size (S), pyramid hierarchy (H), and cutting method (C). $H = 0$ represents the first hierarchy, and the image is cut 0 times. $H = 1$ represents the second hierarchy, and the image is cut 1 time; $H = 2$ represents the third hierarchy, and the image is cut 2 times. Therefore, the number of cutting depends on H. In other words, when $H = h$, the image will be cut h times, and the number of cells generated after cutting is 2^{HC}. Finally, seven different descriptors are obtained in Figure 2, whose size increases exponentially with the number of H and C and has a linear relationship with dictionary size S. The calculation formula of descriptor size η is as follows:

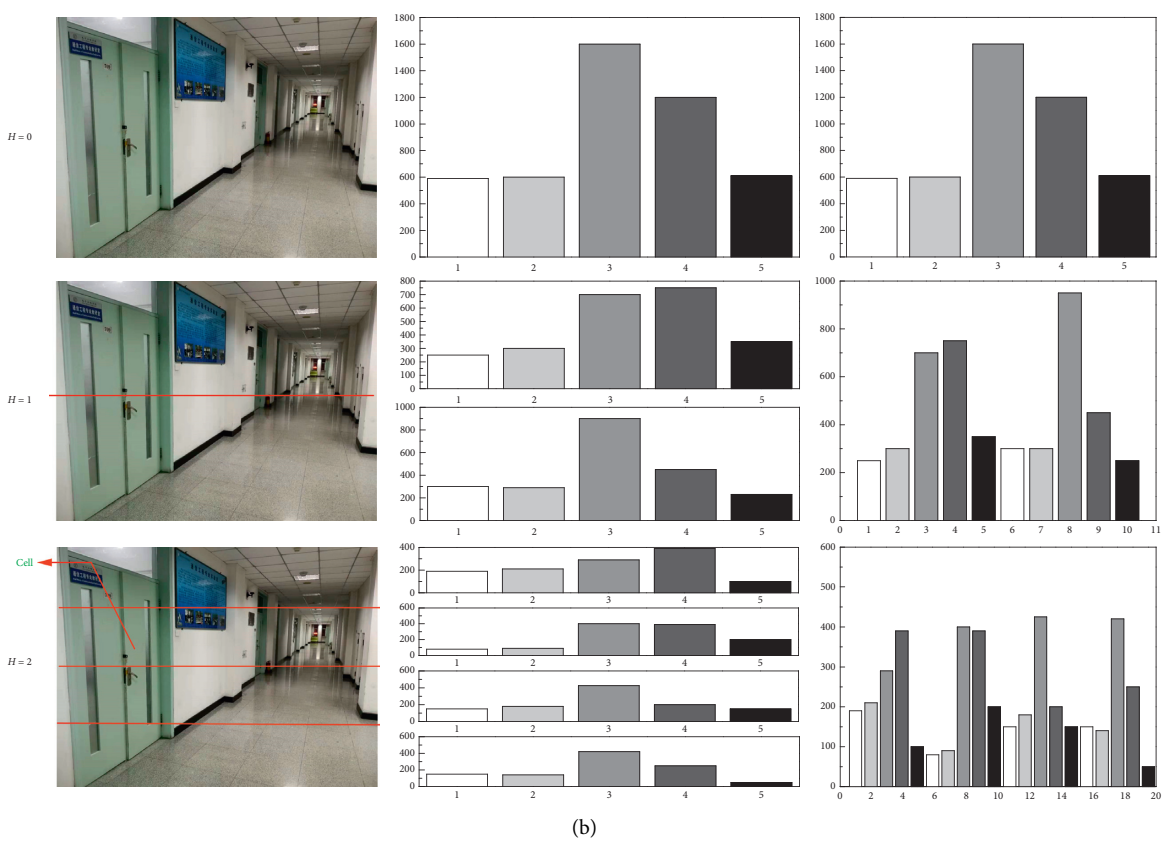
$$\eta = S \cdot 2^{HC}. \tag{3}$$

As we know, image descriptors contain semantic and spatial distribution information of the scene. S will determine the semantic meaning of descriptors, while H and C will focus on the spatial distribution of descriptors, ensuring that more detailed information can be provided. The larger S will provide more detailed semantic information, making features more obvious and more representative. However, if there are a lot of visual words, the histogram will become longer, which will affect the image retrieval and matching process, subsequently. Analogously, a higher pyramid hierarchy contains more detail, while a lower hierarchy is more general.

As can be seen from [12, 13, 38], the standard values of the three parameters are $S = 20$, 50, and 100; $H = 0$, 1, and 2; and $C = 1$ (horizontal and vertical segmentation) and 2 (grid segmentation), respectively. 21 different visual image descriptors and 21 depth descriptors can be obtained by combining these standard values. The reason why the number of descriptors is 21 instead of 27 (3^3) is that $H = 0$ in the pyramid model does not cut the image, with no demands for combination indeed. In other words, for any S, the first pyramid hierarchy will deal with only one descriptor, while the second and third pyramid hierarchies will deal with three descriptors.



(a)



(b)

Figure 2: Continued.

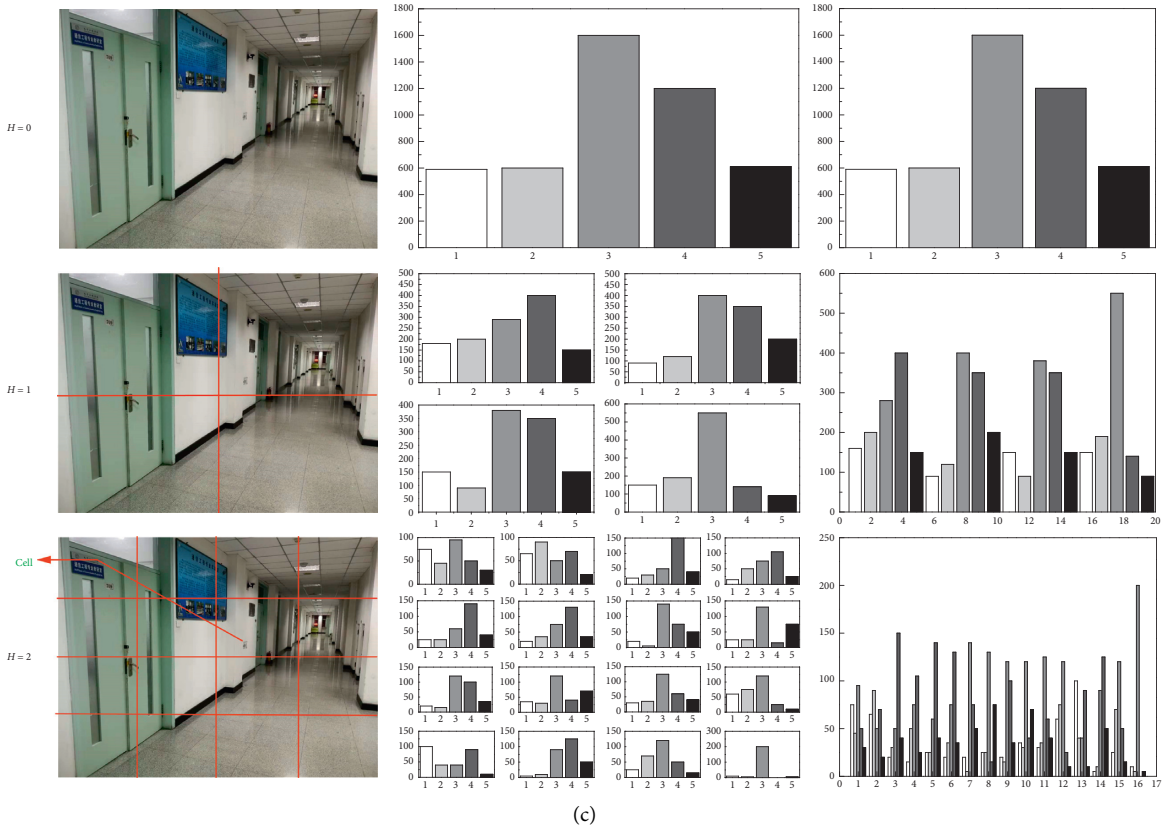


Figure 2: Construction of SPM and generation of the initial descriptor set. (a) Vertical cutting. (b) Horizontal cutting. (c) Grid cutting.

3.3. Descriptors Filtering. In this section, the greedy descriptor filter algorithm (GDFA) will be proposed to find the most valuable descriptors in the initial descriptor set. Since η of the initial descriptors mainly gathered in $(0, 400]$ (as shown in Figure 3), η is divided into three continuous intervals $(0, 150]$, $[150, 350]$,and$[350, \infty)$ for the convenience of descriptor filtering. We assume that large, medium, and small intervals are suitable for our data-gathering platform with small, medium, and high computing power configurations, respectively. The descriptor weight α is related to the descriptor classification accuracy ζ and descriptor size η. In order to obtain descriptors with smaller size and higher accuracy, the calculation formula of the weight α could be defined as follows:

$$\alpha = \frac{\zeta}{\log \eta}. \tag{4}$$

The greedy descriptor filtering algorithm (GDFA) flow is given in Algorithm 1.

At first, the weight of all descriptors is calculated according to equation (4). Next, the descriptor size is divided into $(0, 150]$, $[150, 350]$, and $[350, \infty)$ three continuous intervals, and then the descriptors are sorted in order of weight values from the largest to the smallest. The descriptor with the largest weight in $\mathcal{N}_i$ is filtered and added to the first position in F. If the descriptor weight is greater than 95% of the weight of the previous selected descriptor, that is, $(\alpha_i > 0.95\alpha_{i-1})$, the descriptor is filtered out; otherwise, the next descriptor will be compared. GDFA not only could find out the most valuable descriptors in each interval, but also could filter out descriptors with similar weights.

4. Descriptor Evaluation

4.1. Experimental Database. In order to study the indoor scene classification mechanism, as shown in Figure 4(a), the indoor image data gathering platform with Microsoft Kinect 2.0, independently developed by the laboratory, will be adopted to carry out image data gathering in the Heilongjiang University physical laboratory building. The database contains visual and depth images captured in 9 indoor scenes under different lighting conditions. To cite some examples, Figure 4(b) shows part of the database images.

The database images will be randomly divided into 5 sequences, namely, Training 1, 2, and 3 and Test 1 and 2. The image number for 9 scenes in 5 sequences is listed in Table 1.

4.2. Evaluation Results and Analysis. K-fold cross-validation could be a common accuracy test method, which can effectively avoid over-learning and under-learning. 10-CV (10-fold cross-validation) will be adopted to evaluate the classifier model in this section. To ensure that each cross-validation image is similar, a subset of 30 consecutive images will be randomly assigned to Fold1–Fold10 (represents 10 subsets of the 30 images), which effectively prevented any deviation caused by the time continuity in the data set. Figure 5 shows

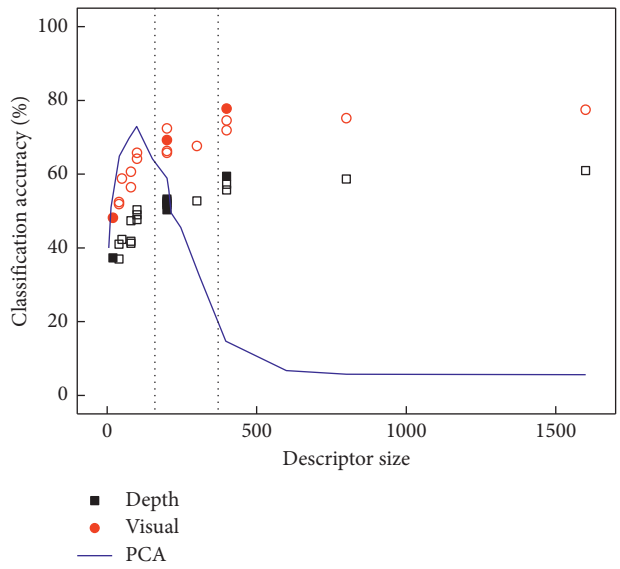


Figure 3: Greedy descriptor filtering algorithm versus PCA.

Input: descriptor list---$\mathscr{L}$
descriptor classification accuracy list---ζ
descriptor size list---η
(1) **for** $j \in \lceil 0, \text{size}(\mathscr{L}) \rceil$ **do**
(2) $\alpha[j] \longleftarrow \zeta[j]/\log(\eta[j])$
(3) **end**
(4) Divide the descriptor size into $(0, 150]$, $[150, 350]$, and $[350, \infty)$ three continuous intervals
(5) **for** $i \in \lceil 1, 2, 3 \rceil$ **do**
(6) Divide $\mathscr{L}$ into new lists $\mathscr{N}_i$
(7) Sort the descriptors in $\mathscr{N}_i$ in order of weight values from largest to smallest
(8) **for** $j \in [0, \text{len}(\mathscr{N}_i)]$ **do**
(9) Filter the descriptor $\mathscr{N}_i$ [1] with the largest weight in $\mathscr{N}_i$ and add $\mathscr{N}_i$ [1] to Φ_i
(10) **if** $\mathscr{N}_i$ $[j-1]$ is filtered and
$\alpha[j] > 0.95 * \alpha[j-1]$ **then**
(11) Add $\mathscr{N}_i$ $[j]$ to α_i
(12) **else**
(13) **end**
(14) **end**
(15) **end**
Output: filtered descriptor list---α

Algorithm 1: Greedy descriptor filtering algorithm.

the distribution of each scene in the data set in each fold of 10-CV and global distribution. It is worth noting that scenes in the data set are not evenly distributed in Fold1–Fold10.

Table 2 shows the classification accuracy of initial descriptors of 42 visual image descriptors and depth image descriptors after 10 times of cross-validation. In SPM, when $H = 0$, for any kind of segmentation type, there is no image cutting and the generated descriptors are identical, so the evaluation results are identical too. By comparing the results of visual images and depth images, we can find that the classification accuracy of depth images is significantly lower than that of visual images. The reason may be that the visual coding technology (visual coding is the mapping between data and visual results) of the depth image is not accurate enough to obtain fine-grained data.

GDFA can find the valuable descriptors from the initial descriptor set, which will facilitate the descriptor combination work in Section 5. Table 3 shows the internal parameters and classification accuracy of the 4 visual image descriptors and 7 depth image descriptors filtered by GDFA, analogously, and the evaluation data are from 10-CV. In other words, the 42 initial descriptors given in Table 2 are reduced to 11 through the filtering of GDFA. These descriptors may have the highest weight in $(0, 150]$, $[150, 350]$, and $[350, \infty)$intervals.

PCA is one of the classical and widely algorithms in current data preprocessing algorithms. Dimensionality

Figure 4: Indoor image data gathering platform (a) and part of database (b).

reduction with PCA can preserve the most important features in high-dimensional data and remove noise and worthless features, which could improve data quality and data processing speed. Figure 3 shows the comparison between the filtering result of GDFA and the dimensional reduction result of PCA (the solid point in Figure 3 is the descriptor obtained by the GDFA, and the dotted line separates three intervals). As observed, when descriptor size is in $(0, 150]$, PCA outperforms both visual descriptors and depth descriptors. But when descriptor size is in $[150, 350]$ and $[350, \infty)$, the performance of PCA begins to decline, which may indicate that GDFA performs better than PCA, especially when the descriptor size is medium or large.

Table 1: The number of images of 9 scenes in 5 sequences.

Scene	Frame				
	Training 1	Training 2	Training 3	Test 1	Test 2
1	438	498	444	511	319
2	140	152	84	95	147
3	119	80	65	109	229
4	421	452	376	392	442
5	408	336	247	307	942
6	664	599	388	692	1287
7	126	79	60	95	223
8	153	96	118	140	193
9	198	240	131	104	241
All	2267	2532	1913	2445	4023

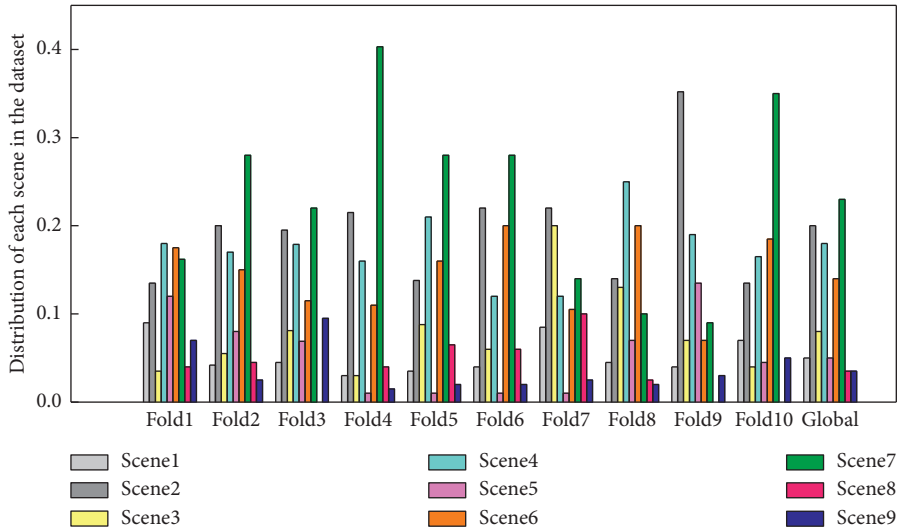


Figure 5: Distribution of each scene in Fold1-Fold10 and the global data set.

Table 2: Evaluation results of initial descriptors.

C	H	Visual			Depth		
		$S = 20$ (%)	$S = 50$ (%)	$S = 100$ (%)	$S = 20$ (%)	$S = 50$ (%)	$S = 100$ (%)
Vertical	0	48.13	58.75	66.20	37.07	42.49	50.06
	1	51.84	63.36	69.53	37.23	48.83	52.55
	2	56.38	65.88	72.09	41.03	50.51	55.44
Horizontal	0						
	1	52.51	64.68	72.01	40.93	47.53	51.78
	2	60.73	72.36	77.81	47.39	53.65	59.40
Grid	0						
	1	56.25	69.02	74.34	41.82	52.95	57.07
	2	67.53	75.26	77.24	52.76	58.37	60.86

5. Descriptor Combination

The most valuable descriptors have been selected by GDFA in Section 4. In order to further obtain the high-quality and highly efficient final descriptor, this section will propose a multiple descriptor combination algorithm (this section only combines two descriptors) although this step might increase the running time of scene classification. There will be two descriptor combination levels, as shown in Figure 6. One is the descriptor level (DL), which can be input to SVM1 after the descriptors of Image1 and Image2 have been connected into one combination descriptor, as shown in Figure 6(a). The other one is the classifier level (CL), which weights the different response results after Image1 and Image2 have been input to SVM1 and SVM2 separately, as shown in Figure 6(b). Also, this section will discuss homologous combinations ($V + V$ or $D + D$) and nonhomologous combinations ($V + D$).

Table 3: Filtering results of GDFA.

Image type	Parameters			Filtering criteria	
	S	H	C	ζ (%)	η (interval)
V1	20	0	—	48.13	20 (1)
V2	50	2	Horizontal	72.36	200 (2)
V3	100	1	Horizontal	72.01	200 (2)
V4	100	2	Horizontal	77.81	400 (3)
D1	20	0	—	37.07	20 (1)
D2	50	2	Horizontal	53.65	200 (2)
D3	50	1	Grid	52.95	200 (2)
D4	100	1	Vertical	52.55	200 (2)
D5	100	1	Horizontal	51.78	200 (2)
D6	50	2	Vertical	50.51	200 (2)
D7	100	2	Horizontal	59.40	400 (3)

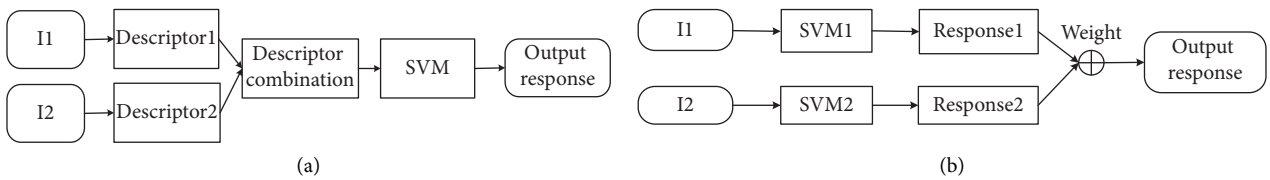


Figure 6: Descriptor combination level. (a) DL. (b) CL.

The combined sequences of training 1, 2, and 3 given in Table 1 will be used as the training set, while Test 1 and Test 2 will be used as the test set. These 5 sequences have the same scene. But it is noted that the light variation in Test 1 is stronger than that in Test 2.

5.1. Homologous Combinations. This section will combine two descriptors extracted from the same image type, namely, V + V or $D + D$, which are called homologous combination. The combination will be carried out at DL and CL, respectively. The test set of SVM could have been composed of two groups of sequences with obvious light differences, Test 1 and Test 2, respectively.

5.1.1. $V + V$. There are 6 different combinations of the 4 depth image descriptors V1, V2, V3, and V4 given in Table 3, which will be applied to DL and CL, respectively. The classification accuracy obtained in Test 1 and Test 2 is shown in Figures 7(a) and 7(b), respectively.

5.1.2. $D + D$. There are 21 different combinations of the 7 depth image descriptors D1, D2, D3, . . ., D7 given in Table 3, which will be applied to DL and CL, respectively. The classification accuracy obtained in Test 1 and Test 2 is shown in Figures 8(a) and 8(b), respectively.

Comparing Figure 7 with Figure 8, we find that the classification accuracy of $D + D$ is generally lower than V + V. The highest classification accuracy in Test 1 and Test 2 achieved by the best depth image descriptor D7 is 48.79% and 65.45%, respectively (while the highest classification accuracy in Test 1 and Test 2 achieved by the best visual image descriptor V4 is 74.76% and 85.78%, respectively). When the best initial descriptor D7 acts as the parent descriptor, the highest classification accuracy of DL is 56.07% in Test 1, while it is 71.86% in Test 2. Apparently, the classification accuracy in Test 2 is still higher than that in Test 1 in $D + D$.

Similar to V + V, DL always outperforms CL in $D + D$. The classification accuracy of combination descriptors in DL is always higher than the parents' descriptors (39 out of 42), while only a few combination descriptors have higher classification accuracy than parents' descriptors in the CL (16 out of 42). The internal parameters of D7 are $S = 100$, $H = 2$, and $C =$ horizontal. $D5$+D7 (56.07%) achieves a favorable effect, and the internal parameters of D5 are $S = 100$, $H = 1$, and $C =$ horizontal. $D2$+D7 (71.86%) also achieves a favorable effect, and the internal parameters of D2 are $S = 50$, $H = 2$, and $C =$ horizontal. The similarity of the optimal combination is $C =$ horizontal, which is verified in Section 4. In addition, the internal parameters of V4 and D7 are $S = 100$, $H = 2$, and $C =$ horizontal. So, we can speculate that high classification accuracy could be obtained by descriptors with such a group of internal parameters, which will be verified in Section 6.

5.2. Nonhomologous Combinations. This section will combine two descriptors extracted from different image types, namely, $V + D$, which is called as nonhomologous combination. There are 28 different combinations of V1, V2, V3, and V4 and D1, D2, D3, . . ., D7 in Table 3, which will be applied to DL and CL, respectively. The specific evaluation process is the same as homologous combination, and the evaluation results are shown in Figure 9.

In Test 2, the highest classification accuracy of CL and DL reaches 80.36% and 92.64%, respectively, while in Test 1, it reaches 72.84% and 81.76%. This is consistent with what

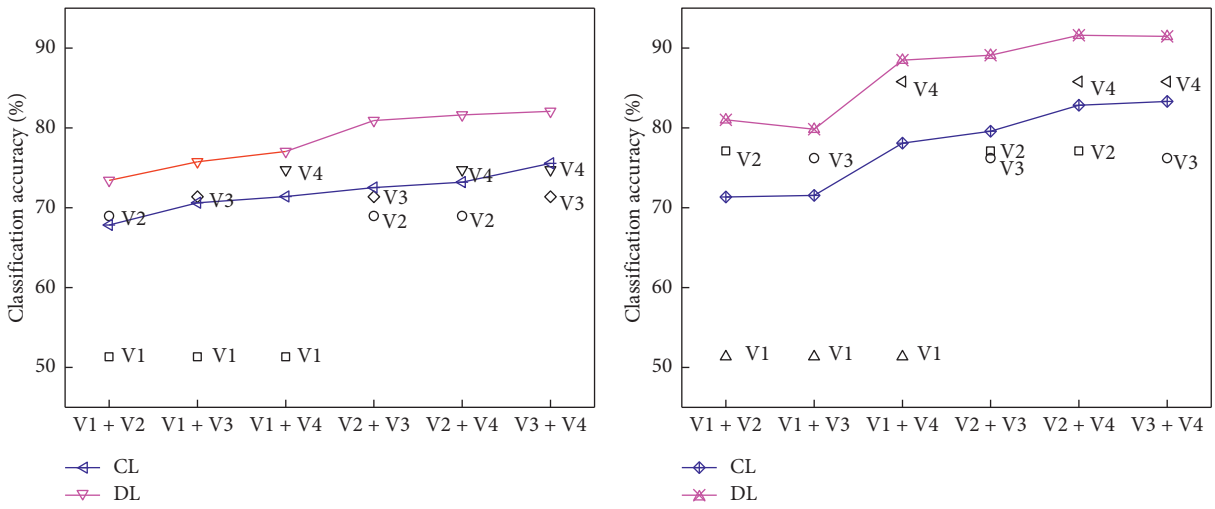


Figure 7: $V + V$ combination evaluation results. (a) Test 1. (b) Test 2.

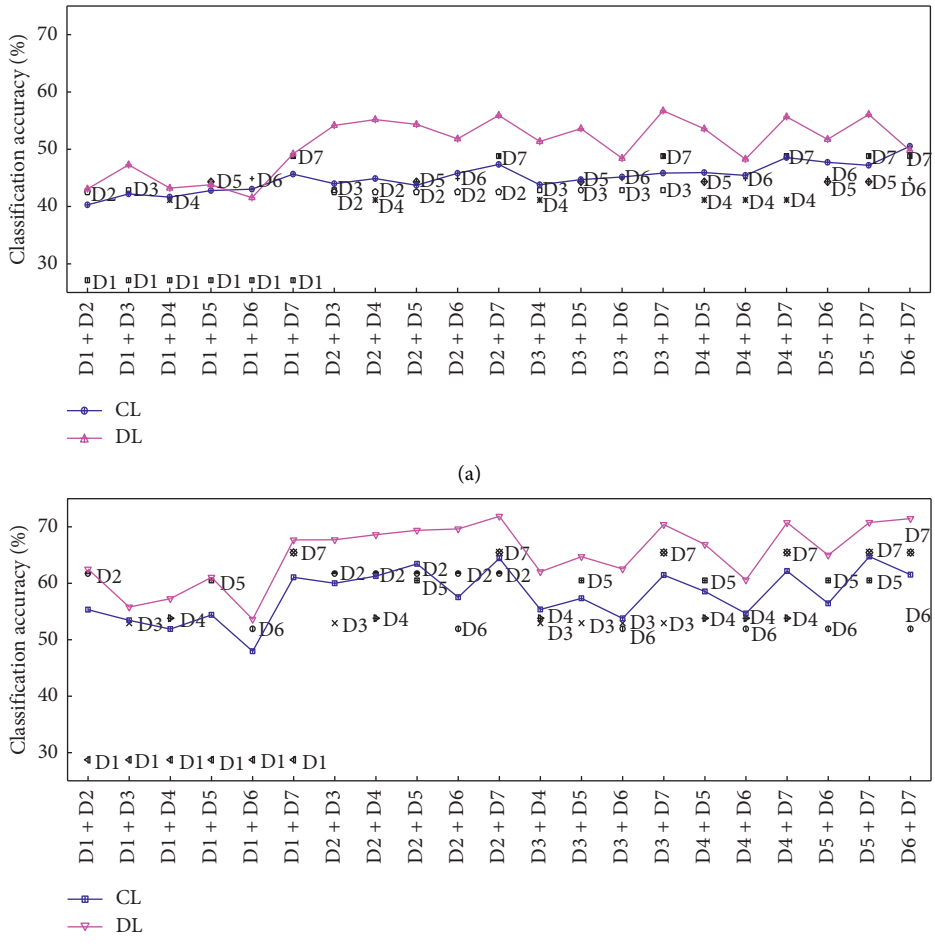


Figure 8: $D + D$ combination evaluation results. (a) Test 1. (b) Test 2.

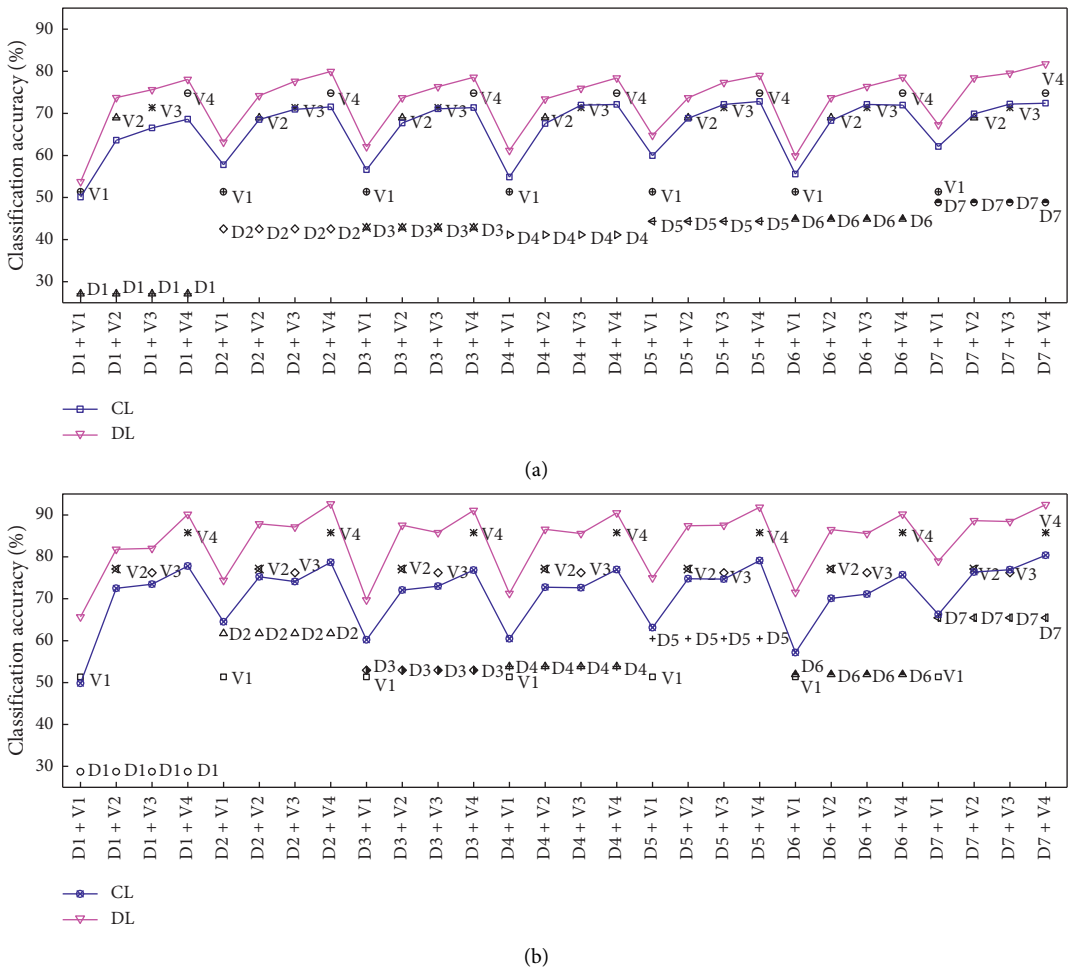


Figure 9: $D + V$ combination evaluation results. (a) Test 1. (b) Test 2.

we found before, the classification accuracy of Test 2 is always higher than Test 1, and DL always outperforms CL.

In CL, the combination with the highest classification accuracy is $D5$+V4 (72.84%) in Test 1. In the meantime, the classification accuracy of V4, which acts a parent descriptor, is 74.76%. The combination with the highest classification accuracy is $D7 + V4$ (80.36%) in Test 2. The classification accuracy of V4, which acts as a parent descriptor, is 85.78%. As shown in Figures 9(a) and 9(b), only a few combination descriptors have higher classification accuracy than parent descriptors in the CL (18 out of 56), the same as in homologous combinations. It shows that the result of CL is not satisfactory.

In DL, the combination with the highest classification accuracy is $D7 + V4$ (81.76%) in Test 1. In the meantime, the classification accuracy of V4, which acts as a parent descriptor, is 74.76%. The combination with the highest classification accuracy is $D7 + V4$ (92.64%) in Test 2. The classification accuracy of V4, which acts as a parent descriptor, is 85.78%. As shown in Figures 9(a) and 9(b), the classification accuracy of combination descriptors in DL is always higher than that in parents' descriptors (56 out of 56).

We can conclude that DL outperforms CL in nonhomologous combination because most combination descriptors in DL outperform their parent descriptor, while the combination descriptors in CL might be difficult to achieve. In addition, no matter in which level, the combinations of the descriptor with excellent performance and the descriptor with poor performance outperform other combinations. To cite some, $D1$+V4 precedes $D1$+V1, $D1$+ V2, and $D1$+V3 in Figure 9(b).

Combining Figures 7–9, we can conclude that the overall effect of $V + V$ and $D + V$ outperforms $D + D$. Sometimes $V + V$ outperforms $D + V$ although nonhomologous combinations contain more comprehensive information. DL combines descriptors before entering a classifier, which may preserve characteristics of the descriptors completely. This may be the reason why DL is always better than CL. So, we only compare the evaluation results of $V + V$ and $V + D$ in DL.

Table 4 lists the best combinations of homologous and nonhomologous in DL, as well as the highest classification accuracy (bold data) obtained in Test 1 and Test 2. The best combination is $V3 + V4$ in Test 1, and the best combination is $D2 + V4$ in Test 2. We recall that the light variation in Test 1 is stronger than that in Test 2. So $V + V$ can be the best in Test 1, while $D + V$ can be the best in Test 2.

Table 4: The best combination.

	V + V	ζ (%)	D + V	ζ (%)
Test 1	**V3 + V4**	**82.09**	$D7$ + V4	81.76
Test 2	V2 + V4	91.60	**D2 + V4**	**92.64**

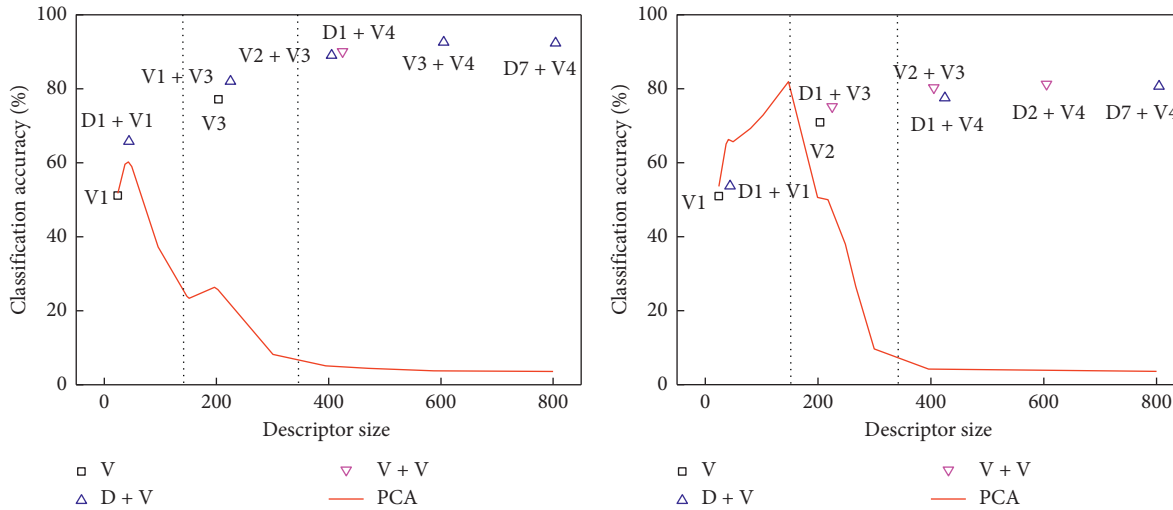


Figure 10: Descriptor size versus classification accuracy and PCA. (a) Test 1. (b) Test 2.

Table 5: Indoor scene classification execution time.

	Step	Parameters	Time (s)
Descriptor generation	Extracting D-SIFT feature	imageSize = 640 * 480	0.0840
	Mapping feature point	$S = 20$	0.0096
		$S = 50$	0.0140
		$S = 100$	0.0218
	Counting histograms	$H = 0$	0.0006
		$H = 1, C = 1$	0.0004
		$H = 2,\ C = 1$ or $H = 1,\ C = 2$	0.0003
		$H = 2,\ C = 2$	0.0002
Descriptor classification	Classifying the input descriptor	$\eta = 20$	0.0010
		$\eta = 50$	0.0016
		$\eta = 100$	0.0029
		$\eta = 200$	0.0062
		$\eta = 400$	0.0131
		$\eta = 800$	0.0291

As shown in Table 3, descriptor size has 8 possible values (including single descriptor or combination descriptor), respectively: 20, 40, 200, 220, 400, 420, 600, and 800. The maximum classification accuracy corresponding to each descriptor size value is compared with PCA results. Figure 10 shows the relationship between classification accuracy and descriptor size in Test 1 and Test 2. As we can see, the classification accuracy of the multiple descriptors fusion mechanism can be improved significantly with the descriptor size from small to middle. Also, the classification accuracy gradually tends to be stable with the descriptor size from middle to large. In Test 1, when descriptor size equals to 400 (large), V2 + V3 (80.94%) gets the highest classification accuracy. In Test 2, when descriptor size equals to 600 (large), $D2$ + V4 (92.64%) gets the highest classification accuracy. PCA achieves high classification accuracy in the condition with small descriptor size. The superiority of the multiple descriptors fusion mechanism becomes obvious with the increasing descriptor size.

5.3. Execution Time. Indoor scene classification is divided into two stages: offline training and online testing. It is assumed that the construction of BoW and classifier training has been completed at the offline stage. Therefore, what affects the

Table 6: Comparison of classification accuracy.

Classification algorithm	ζ (%)
HOG + SVM [28]	77.2
SIFT + SPM [29]	84.2
SIFT + SURF [30]	85.7
Kernel Descriptors + Linear SVM [31]	89.6
Kernel Descriptors + Kernel SVM [31]	90.0
Kernel Descriptors + Random Forest [31]	90.1
Multiple descriptors fusion	**92.6**

running time of the online stage is the generation and classification of descriptors, including 4 steps, as shown in Table 5.

It is worth noting that step 1 adopts imageSize = 640 ∗ 480. Step 2 is related to BoW size (S), so $S = 20$, 50, and 100 are studied, respectively. Step 3 depends on the size and number of image cells, which is related to pyramid hierarchy (H) and cutting method (C). Step 4 is determined by η.

5.4. Algorithm Analysis and Comparison. Under the same database, the classification accuracy obtained by our mechanism will be compared with other fusion methods, as shown in Table 6. The classification accuracy obtained by the algorithms with single feature fusion [28–30] tends to be low for the indoor scene, largely because these algorithms do not filter descriptors. So it seems that the algorithm with single feature fusion is suitable for indoor scene classification. Higher classification accuracy is obtained by the algorithm with multiple features fusion [31], which extracted five different kernel descriptors from the images. After integration, they are trained and classified by Linear SVM, Kernel SVM, and Random Forest, respectively, and obtained 89.6%, 90.0%, and 90.1% accuracy in this experiment. 92.6% accuracy is achieved by our classification mechanism, which has a 2.5% higher value than in [31]. Above all, multiple descriptors fusion mechanism has good performance in indoor scene classification.

6. Conclusion

Aiming at the actual demands for indoor positioning applications, a multiple descriptors fusion model is established and an image classification strategy is proposed to improve the quality and efficiency of descriptors so as to achieve a better indoor scene classification effect. Firstly, the initial descriptor set is formed based on the established SPM. Then, the greedy descriptor filtering algorithm is adopted to select the descriptors with high weight in each descriptor size interval and a valuable descriptor set is obtained. Finally, the multiple descriptors combination algorithm is proposed to obtain high-quality and highly efficient multiple descriptors by combining homologous and nonhomologous images at DL and CL, respectively.

The generation, filtering, and combination of multiple descriptors proposed in this study improve the performance of the classifier. The evaluation results reflect that the multiple descriptors fusion mechanism proposed in this study outperforms the well-known PCA dimensionality reduction technology, especially for the condition with medium or large descriptor size. This strategy not only achieves better results than other feature fusion algorithms, but also solved the limitations of existing scene classification algorithms applied to interior scenes.

Future research will focus on the improvement of the image feature extraction algorithm and the efficiency of constructing visual words by clustering features in the visual BoW model by other clustering algorithms. More attention will be paid to enhance the effectiveness of descriptors when describing image information. At the same time, the improvement of the quality of the depth image will be taken into account so as to make more efficient use of depth data in the process of descriptor filtering and descriptor combination. Alternatively, a more complete data set can be adopted.

Data Availability

The data results used to support the findings of this study are presented in this paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was support by the National High Technology Research and Development Program of China (2012AA120802), National Natural Science Foundation of China (61771186), Postdoctoral Research Project of Heilongjiang Province (LBH-Q15121), University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (UNPYSCT-2017125), and Postgraduate Innovative Research Project of Heilongjiang University (YJSCX2019-166HLJU).

References

[1] P. Nico, P. David, T. Jens et al., "TDoA-based outdoor positioning with tracking algorithm in a public LoRa network," *Wireless Communications and Mobile Computing*, vol. 2018, 9 pages, 2018.

[2] K. H. Lam, C. C. Cheung, W. C. Lee et al., "New RSSI-based LoRa localization algorithms for very noisy outdoor environment," in *Proceedings of the IEEE 42nd Annual Computer Software and Applications Conference*, pp. 794–799, Tokyo, Japan, July 2018.

[3] K. Zhang, S. Chong, Q. Zhou, H. Wang, Q. Gao, and Y. Chen, "A combined GPS UWB and MARG locationing algorithm for indoor and outdoor mixed scenario," *Cluster Computing*, vol. 22, no. S3, pp. 5965–5974, 2019.

[4] X. He, W. Manxing, L. Peng et al., "An RFID indoor positioning algorithm based on support vector regression," *Sensors*, vol. 18, no. 5, pp. 1504–1519, 2018.

[5] X. Yuan, Y. S. Shmaliy, Y. Li et al., "UWB-based indoor human localization with time-delayed data using EFIR filtering," *IEEE Access*, vol. 5, pp. 16676–16683, 2017.

[6] B. G. De, A. Quesada-Arencibia, C. R. Garcia, and J. C. Rodriguez, R. M. Diaz, A protocol-channelbased indoor

positioning performance study for Bluetooth low energy," *IEEE Access*, vol. 6, pp. 33440–33450, 2018.

[7] G. Xiang and Z. Tao, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," *Autonomous Robots*, vol. 41, no. 1, pp. 1–18, 2017.

[8] C. Yujin, C. Ruizhi, L. Mengyun, A. Xiao, D. Wu, and S. Zhao, "Indoor visual positioning aided by CNN-based image retrieval: training-free, 3D modeling-free," *Sensors*, vol. 18, no. 8, pp. 2692–2712, 2018.

[9] X. Aoran, C. Ruizhi, L. Deren, Y. Chen, and D. Wu, "An indoor positioning system based on static objects in large indoor scenes by using smartphone cameras," *Sensors*, vol. 18, no. 7, pp. 2229–2246, 2018.

[10] M. K. Alsmadi, "An efficient similarity measure for content based image retrieval using memetic algorithm," *Egyptian Journal of Basic and Applied Sciences*, vol. 4, no. 2, pp. 112–122, 2017.

[11] M. A. E. Aziz, A. A. Ewees, and A. E. Hassanien, "Multi-objective whale optimization algorithm for content-based image retrieval," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 26135–26172, 2018.

[12] L. Xie, F. Lee, L. Liu et al., "Improved spatial pyramid matching for scene recognition," *Pattern Recognition*, vol. 82, pp. 118–129, 2018.

[13] W. Zhao, H. Luo, J. Peng, and J. Fan, "Spatial pyramid deep hashing for large-scale image retrieval," *Neurocomputing*, vol. 243, pp. 166–173, 2017.

[14] L. Gupta, V. Pathangay, A. Patra et al., "Indoor versus outdoor scene classification using probabilistic neural network," *Eurasip Journal on Applied Signal Processing*, vol. 2007, Article ID 094298, no. 1, p. 123, 2007.

[15] L. T. L. Tao, Y. H. Kim, and Y. T. Kim, "An efficient neural network based indoor-outdoor scene classification algorithm," in *Proceedings of the International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, February 2010.

[16] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2055–2068, 2017.

[17] H. Kebapci, B. Yanikoglu, and G. Unal, "Plant image retrieval using color, shape and texture features," *The Computer Journal*, vol. 54, no. 9, pp. 1475–1490, 2011.

[18] J. K. Patil and R. Kumar, "Analysis of content based image retrieval for plant leaf diseases using color, shape and texture features," *Engineering in Agriculture, Environment and Food*, vol. 10, no. 2, pp. 69–78, 2017.

[19] A. Raza, T. Nawaz, H. Dawood, and H. Dawood, "Square texton histogram features for image retrieval," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2719–2746, 2019.

[20] W. Tahir, A. Majeed, and T. Rehman, "Indoor/outdoor image classification using GIST image features and neural network classifiers," in *Proceedings of the International Conference on High-Capacity Optical Networks & Enabling/emerging Technologies (HONET)*, Islamabad, Pakistan, December 2015.

[21] L. Ju, K. Xie, H. Zheng, B. Zhang, and W. Yang, "GPCA-SIFT: a new local feature descriptor for scene image classification," *Communications in Computer and Information Science*, vol. 663, no. 4, pp. 286–295, 2016.

[22] J. Shi, H. Zhu, J. Wang et al., "Indoor scene classification algorithm based on information enhancement of vision sensitive area," *Moshi Shibie Yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence*, vol. 30, no. 6, pp. 520–529, 2017.

[23] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *2013 Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1417–1424, Sydney, Australia, March 2014.

[24] Y. Zheng, J. Pu, H. Wang et al., "Indoor scene classification by incorporating predicted depth descriptor," *Pacific Rim Conference on Multimedia*, vol. 10736, pp. 13–23, May 2018.

[25] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, pp. 719–723, Nice, France, December 2015.

[26] A. Janoch, S. Karayev, Y. Jia et al., "A category-level 3D object dataset: putting the Kinect to work," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1168–1174, Barcelona, Spain, November 2011.

[27] N. Silberman, D. Hoiem, P. Kohli et al., "Indoor segmentation and support inference from RGBD images," in *Proceedings of the 12th European conference on Computer Vision (ECCV)*, pp. 746–760, Springer, Berlin, Germany, October 2012.

[28] L. Jin, L. Quan, and A. Qingsong, "Research of image classification based on fusion of SURF and global feature," *Computer Engineering & Applications*, vol. 49, no. 17, pp. 174–177, 2012.

[29] R. Rani, S. Kumar Grewal, and K. Panwar, "Object recognition: performance evaluation using SIFT and SURF," *International Journal of Computer Applications*, vol. 75, no. 3, pp. 39–47, 2013.

[30] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots & Systems*, pp. 821–826, San Francisco, CA, USA, December 2011.

[31] K. Bregar, M. Mohorcic, and Y. Yang, "Improving indoor localization using convolutional neural networks on computationally restricted devices," *IEEE Access*, vol. 6, pp. 17429–17441, 2018.

[32] Y. Zhou, Y. Zhou, Q. Liu et al., "Research on a DSIFT algorithm applicable to image mosaicking," *Journal of Xian Jiaotong University*, vol. 49, no. 9, pp. 84–90, 2015.

[33] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: two decades of texture representation for texture classification," *International Journal of Computer Vision*, vol. 127, no. 1, pp. 74–109, 2019.

[34] L. Lifeng, M. Yan, Z. Xiangfen, Y. Zhang, and S. Li, "High discriminative SIFT feature and feature pair selection to improve the bag of visual words model," *Iet Image Processing*, vol. 11, no. 11, pp. 994–1001, 2017.

[35] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.

[36] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Systems with Applications*, vol. 67, pp. 12–18, 2017.

[37] E. Lee, M. Schmidt, and J. Wright, "Improved and simplified inapproximability for k-means," *Information Processing Letters*, vol. 120, pp. 40–43, 2017.

[38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, October 2006.