


Article

# Scene Semantic Recognition Based on Probability Topic Model

Jiangfan Feng and Amin Fu \* 

College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; fengjf@cqupt.edu.cn

\* Correspondence: fam15539278227@163.com

Received: 4 February 2018; Accepted: 15 April 2018; Published: 19 April 2018



**Abstract:** In recent years, scene semantic recognition has become the most exciting and fastest growing research topic. Lots of scene semantic analysis methods thus have been proposed for better scene content interpretation. By using latent Dirichlet allocation (LDA) to deduce the effective topic features, the accuracy of image semantic recognition has been significantly improved. Besides, the method of extracting deep features by layer-by-layer iterative computation using convolutional neural networks (CNNs) has achieved great success in image recognition. The paper proposes a method called DF-LDA, which is a hybrid supervised–unsupervised method combined CNNs with LDA to extract image topics. This method uses CNNs to explore visual features that are more suitable for scene images, and group the features of salient semantics into visual topics through topic models. In contrast to the LDA as a tool for simply extracting image semantics, our approach achieves better performance on three datasets that contain various scene categories.

**Keywords:** scene semantic recognition; convolutional neural networks; DF-LDA

## 1. Introduction

As one of the most basic and important forms of multimedia information, images have been widely used in the fields of image classification, target detection, geographical annotation, and so on, because of its intuitive appearance and rich content. How to effectively manage the surge of image data and obtain semantic information from these data has become an urgent task. In recent years, the model of latent Dirichlet allocation [1] has been successfully and rapidly applied to many fields by processing the texts or images to obtain thematic variables, and use them as a basis of classification or other processing, such as processing of texts [2], image retrieval [3], remote sensing images [4], data mining [5], and so on. However, the LDA model has higher requirements for image data, and the change of brightness, illumination, and scale of images will bring great difficulty to image recognition.

In order to overcome these shortcomings, the work in this paper considers the use of a deep network model to extract the features of an image, and deduces the topic distribution in combination with LDA. Based on these motivations, a DF-LDA strategy is proposed. Deep learning is a learning algorithm that simulates human brain thinking, of which the most important application is convolutional neural networks [6] for image classification. CNNs is mainly used to identify two-dimensional graphics of displacement, scaling, and other forms of distortion invariance, which can adapt to the influence of complex and changeable images on the extracted image features in practical application scenarios. The network avoids the complex pre-processing of images and can directly input the original images, which removes the subjective influence of traditional manual feature extraction on the images. In addition, this paper also takes advantage of LDA's powerful ability to deduce the topics distribution. The use of CNNs helps us find more suitable deep features to adapt LDA more adaptively. Based on latent Dirichlet allocation of deep features, a strategy called DF-LDA effectively

reveals the intrinsic properties of the original data. This method of representing an image using the topic hybrid vector instead of the visual feature is more in line with the human perception process. It crosses the semantic gap from feature extraction directly to scene semantics. Experimental results show that the proposed method has good feature representation ability for real scene datasets and has good performance for scene semantic recognition.

## 2. Related Works

With the explosive growth of data volumes, there are many methods for visual object recognition. Lowe [7] has proposed an image local feature descriptor SIFT based on scale space, with the invariance of image scaling, rotation, and affine transformation. It makes the image classification effect better in outdoor scene due to the good discriminating ability. Quattoni and Torralba [8] proposed an indoor scene description method combining Gist features and ROI (region of interest). However, ROI acquisition in region of interest not only requires a lot of computation time, but also needs image segmentation before obtaining ROI regions. This method of classifying images by low-level features is unsatisfactory in terms of classification accuracy. The topic model constructed by latent variables can capture the structural distribution of image expression and improve the image classification accuracy. This method not only considers the image as a collection of features, but also contains rich image information, which is close to the way humans see the image. Typical topic models are the probabilistic latent semantic analysis (pLSA) [9] model proposed by Hofmann et al. and latent Dirichlet allocation (LDA) model mentioned by Blei et al. Bosch [10] used four different types of visual features in combination with pLSA for scene classification. Fei-Fei and Perona [11] used the pixel-intensity and SIFT feature fusion to obtain the bag of words, and used LDA to classify the scenes.

In the 2012 ILSVRC competition, Hinton et al. used deep learning model called AlexNet to classify the millions of dataset ImageNet, which outperformed traditional classification methods. In the training phase, the deep neural network model uses the original natural image without any human influence. It takes the end-to-end processing and puts the image preprocessing and feature extraction into a black box, which can learn the rotation, twisting, etc. effectively image representation of invariance features. CNNs has excellent performance in solving the general rough classification problem, but the recognition rate of semantic analysis is not enough for the content rich and delicate scenes. It has become a trend to combine convolutional neural networks with traditional classification methods to improve the recognition rate of scene images. Niu et al. [12] combined convolutional neural networks with support vector machines (SVM) and experimented with the MNIST database, and proved that this hybrid approach can further improve the classification accuracy of digital images. Zhen et al. [13] proposed a method to visualize the topic features of untagged datasets using a combination of a pre-trained AlexNet model with a potential Dirichlet assignment (LDA). Wang et al. in 2016 [14] proposed three ways to enhance VLAD coding and aggregated these methods to further improve recognition performance. In 2017, they [15] designed a new network called PatchNet to simulate local patches and proposed a new VSAD coding method to aggregate the features of each local block. The combination of the two methods can achieve excellent results in the scene recognition tasks. Wang et al. [16] studied the relationship between objects, scenes, and events. Inspired by this, they proposed an iterative selection method to find the most relevant subset of representations in objects and scenes. Furthermore, they developed three transfer methods, which exploit a multi-task learning network for event CNNs. It can absorb knowledge from other networks and datasets as a priori knowledge of the network. In this way, the generalization ability of the network is improved, and it is predominant for event datasets. These methods all significantly improve the performance of visual recognition tasks, but they do not consider how to find higher-level semantic information based on visual features, so that the expression of the image is more concise. BOT was proposed as a topic model [17]. The method that replaces the visual features with the topics can reduce the feature dimension and mine higher-level semantic information. However, the method is insufficient to consider the complexity and variability of scene images. Generally speaking, the DF-LDA algorithm

proposed in this paper delivers higher-level topic semantic information by abstracting latent relevance of visual features. This achieves excellent performance in fine-grained scene semantic recognition.

### 3. Proposed Approach

#### 3.1. Overview

The summary of the scene semantic recognition method in this paper is shown in Figure 1. The work in this paper is to develop a hierarchical feature learning structure that combines feature extractor and unsupervised clustering to improve classification performance. Considering the ability of the convolutional neural network to learn the varied images, we chose it as a feature extractor for extracting scene images. In order to mine the potential variables of the image, we chose LDA successfully applied in the text area to model the extracted features and project the original feature space into the topic space. In this way, each image is represented as a mixed distribution of multiple potential topics, which is similar to the histogram of regions represented by all topics. The DF-LDA method proposed in this paper combines the advantages of these two models so that the network can be more beneficial to solve the problem of scene recognition. In the following sections, the work will elaborate on the details of the algorithm.

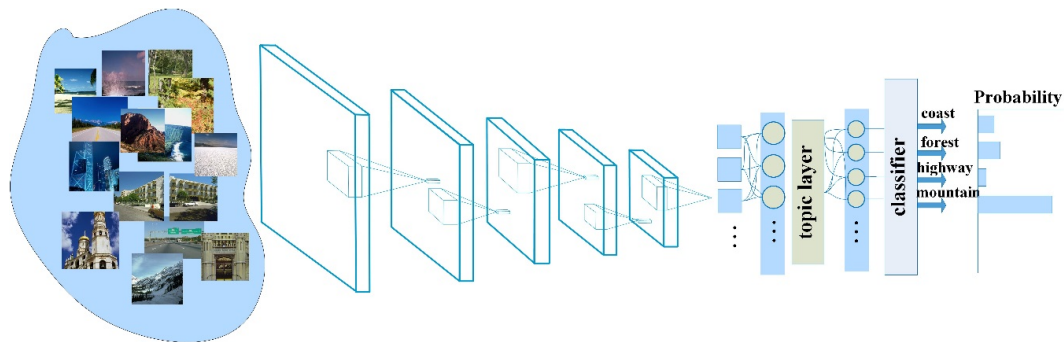


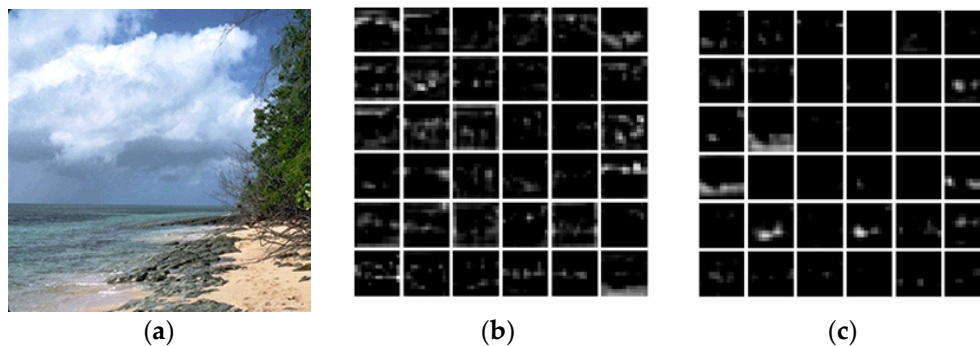
Figure 1. The architecture of our model.

#### 3.2. Deep Feature Extraction and Representation

The image feature extraction algorithm provides direct information for mining scene semantics, so the feature extracted quality will directly affect the correctness of semantic recognition. Razavian et al. [18] proved that the features extracted using convolutional neural networks have excellent effects in image classification and retrieval. Considering the fact that scene semantic recognition is often affected by many conditions, we choose a powerful convolutional neural networks model tool for image tasks to extract image features. The feature detector of CNNs learns images features from the training data implicitly, which avoids the influence of other unfavorable factors. Furthermore, the network reduces the complexity of the network with its special structure of local connection and weight sharing. In particular, the feature that the multi-dimensional input vector image can be input directly into the network avoids the complexity of data reconstruction during feature extraction and classification.

This paper uses the AlexNet model (pre-training at ImageNet) to train our own datasets, separating datasets by a ratio of 7:3, 70% of them as training sets and 30% as testing sets. Insufficient data sets may result in over-fitting, and we reduce the limit by performing data augmentation. Finally, the image input to the network is set to 227 by 227. This paper studies how to automatically extract the features of the scene image, in order to find the features that are more suitable for embedded topic model, and achieve efficient recognition of the scene images, we have visualized the feature maps of learning. By observing the feature map, we can intuitively see that the front feature map is more complex, indicating that the extracted features are relatively simple, and the

feature map of the next layer is more abstract and more prominent than the features of the previous feature map. Figure 2 shows the characteristics of the third and fifth layers in the network.



**Figure 2.** Visualization of feature maps in the network. (a) original image; (b) conv3; (c) conv5.

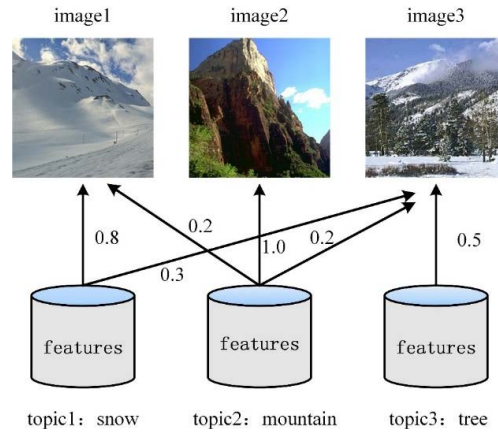
In order to verify the results of the visualization, we try to embed our method behind each convolutional layer of the network for comparison. Repeated experiments prove that the output of the fifth convolution layer is more unique and completely discernible. Therefore, this paper uses the parameters trained on ImageNet dataset as initialization parameters, and on this basis, the fifth convolution layer is re-trained to get fine-tuning parameters, and get the required feature extractor. The output of the feature extractor is an  $nC$ -dimensional vector,  $n$  is the number of convolution kernels, and  $C$  is the number of blocks in the convolution layer. Qiao et al. [19] found that scene classification has a high correlation with local semantics, and that middle-level semantics produced by convolutional features can further improve classification performance. For a particular image  $w_d$ , the LDA takes the frequency of visual features as input. The traditional frequency vector for the visual features of the LDA input was obtained by k-means. The initial cluster center selection of this algorithm is random, which makes iterative calculation more complicated and negatively affects the scene recognition. In this paper, the DF-LDA model uses the k-mean++ [20] clustering algorithm to generate a visual dictionary. Compared with the randomness of the initial clustering centers selected by the traditional LDA model, we define the vector of features as  $F = \{f_1, f_2, \dots, f_C\}$  and choose one of the feature vectors  $f_i$ , calculate the Euclidean distance  $d_i$  between the feature vector and the remaining feature vectors, and select the vector farthest from the feature European vector as the initial cluster center. Continuing this process until you select  $N$  initial cluster centers. The purposeful selection of clustering centers can greatly speed up the convergence of clustering centers and improve the efficiency of the algorithm. By mapping the local features of the image to the corresponding words on the visual dictionary, a histogram of the number of appearances of each visual feature on the image can be obtained, and the image is represented by the appearance frequency of each visual word.

### 3.3. Topic Model

#### 3.3.1. Topic Modeling for Image Representation

Topic models are the link between visual features and images, and contain rich high-level semantic information. In order to excavate the topic semantics in the image, this paper uses LDA to project the original feature space into the topic space. Figure 3 is an example of a topic model. There are three images of a mountain in the figure, and each image can be described with dozens or more visual features. Here, the feature refers to the frequency vector of features in the visual dictionary. Because it is the input data for LDA, this paper also refers to this frequency vector as a feature. We abstract three topics “snow, mountain, and tree” from images and combine the probability distribution of each topic in the image, then images can be described with a mixture of topics. The decimal on the arrow in the figure indicates the probability of different topics for each image, then the topic distribution of the

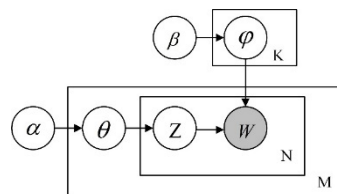
three images can be expressed by (0.8, 0.2, 0), (0, 1.0, 0), (0.3, 0.2, 0.5). This example shows that using the topic model to represent the image as a form of probability distribution of multiple topics can maintain a significant difference between the images, which is an efficacious way to deal with scene recognition problems.



**Figure 3.** The schematic of the topic representation of the images.

### 3.3.2. Topic Model

Topic model is a modeling method of the implicit topic of visual words. Through the deduction of the image topic, the hidden visual topic can be obtained. The LDA model is the best probabilistic topic model for statistical text classification [21], so we model the features of the image as words of the text and assume that there are hidden topic variables between the feature and the image. Different from the traditional topic model based on low-level features, we propose DF-LDA, which means that the experiment uses deeper feature training LDA model. Figure 4 shows a graph model representation of the LDA model used for image processing. Given the image set with  $M$  images  $D = \{w_1, w_2, \dots, w_M\}$ , according to the procedure in Section 3.2, each image can be described as  $N$  clusters of label sequences, denoted by  $q = (q_1, q_2, \dots, q_N)$ . For the LDA generation process, the topic  $Z_{d,n}$  of the feature  $w_{d,n}$  in image  $w_d$  is generated by the multinomial distribution of the parameter  $\theta_d$ .  $\alpha$  and  $\beta$  are parameters of Dirichlet distribution.  $\theta_d$  and  $\varphi_{Z_{d,n}}$  are used to generate multinomial distribution of topics for one image and multiple distributions of visual features for each topic.



**Figure 4.** LDA graph model representation.

According to the LDA graph model, the joint distribution of all the variables is shown in Equation (1)

$$p(W_d, Z_d, \theta_d, \Phi | \alpha, \beta) = \prod_{n=1}^{N_d} p(w_{d,n} | \varphi_{Z_{d,n}}) p(Z_{d,n} | \theta_d) p(\theta_d | \alpha) p(\Phi | \beta) \quad (1)$$

Because  $\alpha$  produces topic distribution  $\theta$ , which identifies the specific topic, and  $\beta$  produces word distribution  $\varphi$ , which identifies the specific word. Therefore, Equation (1) is equivalent to the joint probability distribution  $p$  of all the variables expressed by Equation (2).



$$p(W, Z|\alpha, \beta) = p(W|Z, \beta)p(Z|\alpha) \quad (2)$$

Our goal is to calculate the probability that the  $n$ th feature of the  $d$ th image is assigned to the topic  $Z$ , i.e., the posterior distribution of  $p(Z|W)$  (3). This distribution cannot be directly calculated, because the sum of the denominators cannot be decomposed factor, so the Gibbs sampling method is chosen to estimate and infer the hidden variables of LDA model. The Gibbs sampler of this model samples the topic  $Z$  of each visual feature  $W$ , avoiding the estimation of the actual parameters  $\theta$  and  $\varphi$  by integration. As long as the subject of each visual feature is identified, the values of  $\theta$  and  $\varphi$  can be calculated from frequency statistics. After sampling, the distribution  $\theta_{d,k}$  of the topic  $k$  in the image  $d$  and the distribution  $\varphi_{k,t}$  of the feature  $t$  in the topic  $k$  can be calculated using Equations (4) and (5), respectively.

$$p(Z|W) = \frac{P(W, Z)}{\sum_Z p(W, Z)} \quad (3)$$

$$\theta_{d,k} = \frac{n_d^k + \alpha_k}{\sum_{k=1}^K n_d^k + \alpha_k} \quad (4)$$

$$\varphi_{k,t} = \frac{n_k^t + \beta_t}{\sum_{t=1}^V n_k^t + \beta_t} \quad (5)$$

We propose a more efficient method of scene image representation, DF-LDA, which is a probabilistic topic model based on depth feature representation and latent Dirichlet assignment. The method learns visual words from the output of the last convolutional layer using an enhanced version of the clustering algorithm. In other words, the values of all convolution kernels are considered instead of only the maximum, and the method can show the characteristic of the image through the co-occurrence of visual features. If the topic represents a certain area in the image, the topic distribution of the image is a histogram representation of the statistical topic. Figure 5 shows an overview of the scene semantic recognition model based on the DF-LDA strategy.

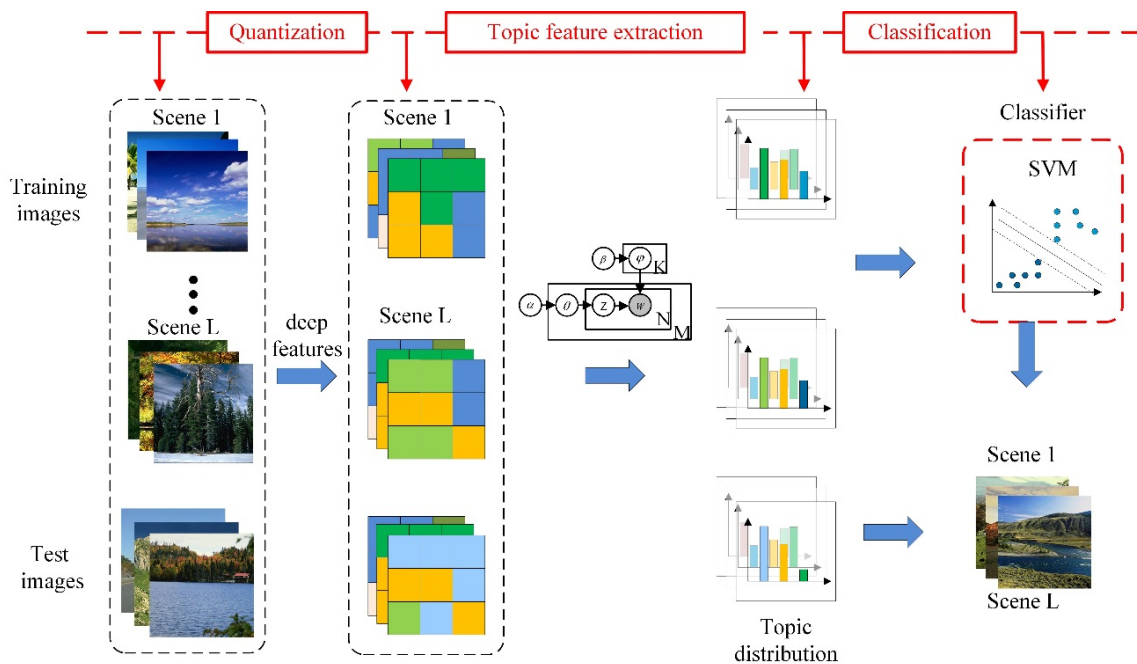


Figure 5. Overview of DF-LDA model in scene classification.

## 4. Experiments and Results

In order to verify the effectiveness of the proposed method, we first describe the datasets used in the experiment, then describe the influence of the visual dictionary capacity and the number of topics on scene recognition, and finally discuss the results of scene semantic recognition.

### 4.1. Description of Datasets

The purpose of this paper is to construct a generic network model that identifies the semantics of the scene, we use three different types of datasets to evaluate results of the experiment. We use three different types of datasets to evaluate the results of the experiment, two of which are benchmark datasets, and one is a self-built dataset. Figure 6 shows some sample images of the self-built dataset, this is a more challenging dataset as can be seen from the Figure 6. We describe these three datasets as follows.

UIUC Sports is a sports event dataset provided by Fei-Fei [22]. This dataset has 1579 images and can be divided into 8 complex motion behavior categories, such as bocce, croquet, polo, etc. Image data is downloaded from the Internet, so the image content is very noisy. In each scenario, image content includes changes in scale, clothing changes, lighting changes, background, and others. The number of images in the category is uneven, but all are .jpg format.

Labelme is a natural scene dataset named Labelme provided by [23]. The dataset is from Oliva & Torralba Image Library. It contains eight natural scene image categories with a total of 2688 images, abbreviated as OT8. Images are sampled from different perspectives, different spatial patterns, different seasons, different rotation angles, etc.

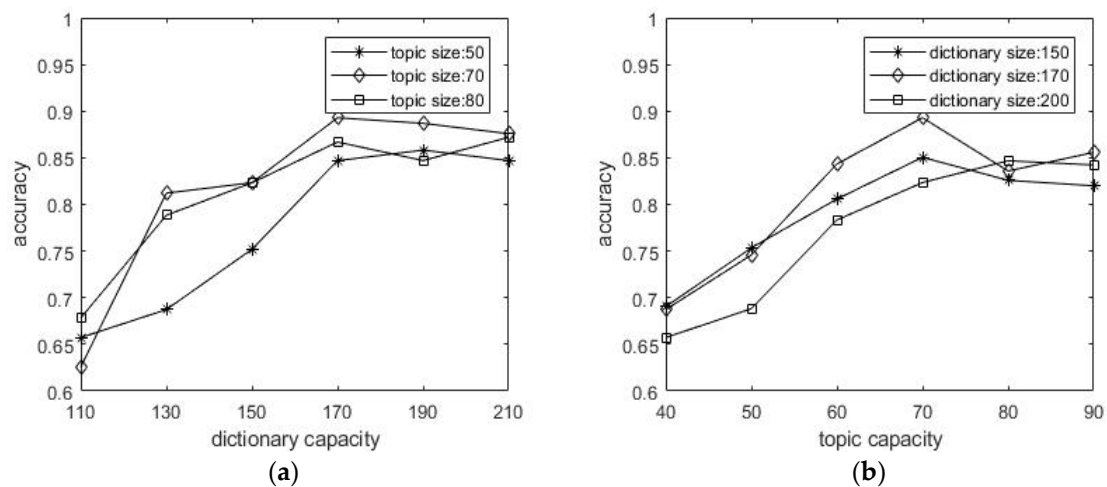
The combined dataset is a self-constructed dataset containing 10 classes. The data is selected from the two datasets already introduced and the indoor67 [23] with only slight internal changes. This dataset contains images such as “mall”, “tallbuilding”, “bowling”, etc. It can be seen from the sample image in Figure 4 that the dataset contains the unique complexity of scene images such as in-class inconsistency, inter-class consistency, light variability, and so on. This is a more challenging dataset that basically includes a variety of comprehensive factors that hinder scene recognition.



Figure 6. A few examples in the combined dataset.

#### 4.2. Parameter Analysis

In the experiment, the dictionary capacity and the number of topics are two very important parameters that affect the experimental results. This paper discusses the relationship between the dictionary capacity and the number of topics on the Labelme dataset. The size of the cluster center  $N$  needs to be set in advance. We set the number of image themes for the case of (50, 70, 80), and the dictionary capacity respectively (110, 130, 150, 170, 190, 210), analyze the impact of dictionary capacity on the classification accuracy. Similar to the method of determining the number of lexical capacity, we have a fixed dictionary capacity (150, 170, 200), and topicset the number of the topics (40, 50, 60, 70, 80, 90), and analyze the influence of the number of topics on classification accuracy. Figure 7 shows that the optimal selection of 170 dictionary sizes and 70 topics can be achieved in the experiment.



**Figure 7.** The Influence of dictionary capacity and topic capacity on the recognition accuracy. (a) Fixed the topic size and analyzed the influence of different dictionary capacity on the recognition accuracy. (b) Fixed the dictionary size and analyzed the influence of different topic capacity on the recognition accuracy.

#### 4.3. Results of Different Classifiers

This paper compares two commonly used classifiers in scene semantic recognition methods, which are KNN and SVM. According to the preprocessing steps, feature extraction, clustering, and topic modeling, KNN and SVM are used as classifiers in the two different datasets. The results show that using SVM as a classifier is better than KNN, see Table 1. So in the experiment using SVM as a classification of learning.

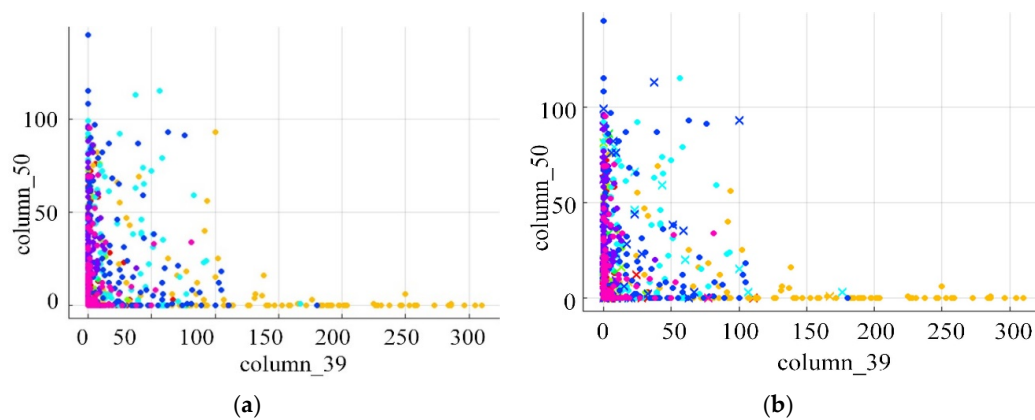
**Table 1.** Recognition accuracy of different classifiers.

	UIUC Sports	Labelme
KNN	82.36%	85.58%
SVM	87.34%	91.04%

We randomly selected two-dimensional data on the Labelme dataset for display. Each image is essentially a multi-dimensional column vector before it is sent to the classifier, so the following figure shows the data information of different dimensions of the column vector. In Figure 8, different colors represent different categories, (a) is the corresponding scatter point of the DF-LDA vector in the selected dimension, (b) is the classification results. In the Figure 8b, the dots indicate the data that is correctly classified and the data that is misclassified as a cross. As can be seen from the figure, most of



the data points can be assigned to the correct category, there are still a small number of data points cannot be correctly classified.



**Figure 8.** Display any two-dimensional classification results in Labelme. (a) Image data point display before entering the classifier. (b) Display of image prediction results in classifier.

#### 4.4. Results and Discussion

According to the method described in the previous section, in this section we will report the performance of the experiment on UIUC Sports and Labelme compared with previous research. As shown in Table 2, in order to verify the effectiveness of the proposed method for scene semantic recognition more comprehensively, we compared the best experimental results with the best results of the other most advanced methods on different data sets.

**Table 2.** The classification accuracy of different methods on different datasets.

No.	Method	UIUC Sports (%)	Lamelme (%)
1.	Topic Feature [24]	70.63	80.75
2.	OB [25]	82.30	86.35
3.	scLDA [26]	81.60	–
4.	LScSPM [27]	85.31 ± 0.51	–
5.	CNN + LDA	85.58	88.43
6.	Our approach	87.34	91.04

The best results for this method in the datasets for UIUC Sports and Labelme were 87.34% and 91.04%, respectively. In the UIUC Sports dataset, the recognition rate of the best method is about 85.31% [27]. It proposes a Laplacian sparse coding method, which utilizes the dependence between local features to improve the recognition accuracy. The best method on the dataset Labelme has an accuracy of 86.50% [25], which uses threshold filter pooling to reduce noise accumulation in the histogram and uses the Matthew effect normalization method to highlight useful information. However, most of these methods preprocess the image, such as filtering and histogram equalization. Although this plays an important role in enhancing the image content, it also filters out some of the details. In addition, these methods are based on manual feature extraction, which is time-consuming and labor-intensive, and the influence of subjective factors is also the main factor that causes the image recognition to not reach the optimum. Since deep neural networks can learn good features automatically from the original image, we try to get the image semantic by combining the network with the LDA. This method can improve the recognition accuracy to a certain extent. However, the traditional clustering algorithm has randomness to the initial clustering centers, which leads to the instability of the recognition results. This paper chooses the farthest feature between the Euclidean distances as the initial clustering center to limit this shortcoming. The approach in this

paper makes use of the power of LDA to better infer the implicit topic of images, with the best feature extractor showing the best performance on the two datasets involved.

In this paper, we chose the confusion matrix that represents the accuracy evaluation to visualize the experimental results. Figures 9 and 10 shows the visualization of the confusion matrix on two different datasets. The value of each column of the confusion matrix is equal to the number of actually measured images in the classification image corresponding to the corresponding category, and the value in each row represents the number of corresponding categories of test images.  $C_{ij}$  in the confusion matrix  $C$  indicates the number of images divided into scene  $j$  belonging to scene  $i$ . The number of diagonals is the number of images correctly classified, the classification accuracy is the sum of the diagonal divided by the total number of images. Therefore, the value of the diagonal is much larger than other values, indicating that the classification accuracy is high, and the classification performance is good.

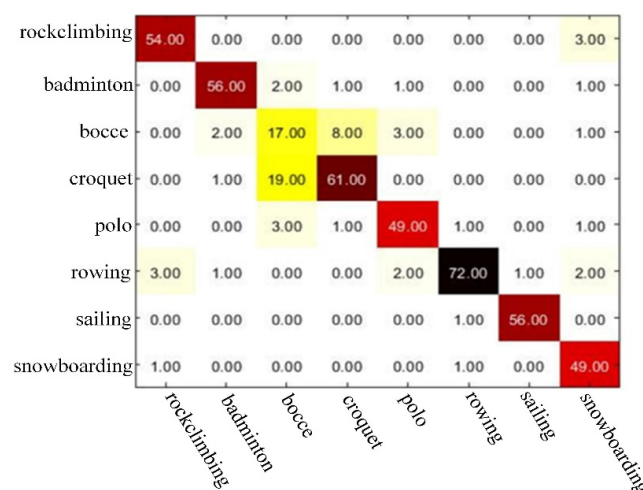


Figure 9. The visual predicted matrix of UIUC Sports.

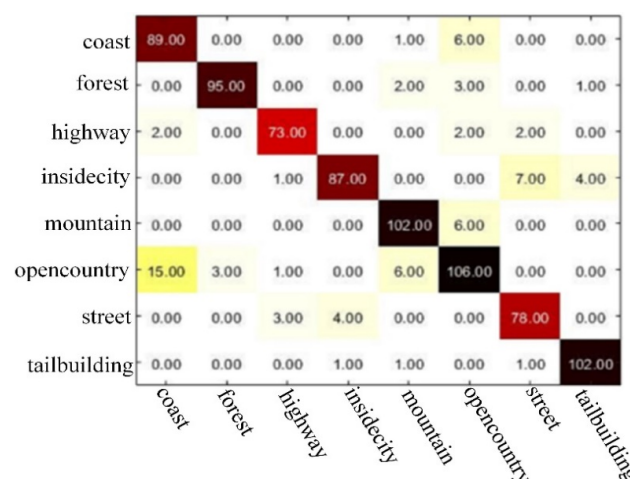


Figure 10. The visual predicted matrix of Labelme.

In order to evaluate the method proposed in this paper more objectively, we chose the other four methods to conduct comparative experiments on the same dataset. There is a traditional method of scene recognition. In addition, a topic model based on an enhanced clustering algorithm is reproduced in this method. We have also fine-tuned a deeper convolutional neural network with a topic model after its last convolutional layer. The third method is based on the method of this paper, but no

enhanced clustering algorithm is used. In order to more equitably evaluate the performance of the method, the settings of all parameter values are the same in the experiment. Table 3 is the average of five experiments for each method. As can be seen from the data in the table, the method in this paper is reasonable. However, if the experiment uses a deeper network as the feature extractor, the effect of the experimental effect will be slightly higher than the proposed method in this paper, but this is premised on the large amount of time and cost.

**Table 3.** The classification accuracy of different methods from the combined datasets.

No.	Method	Combined (%)
1.	SIFT + LDA	65.83
2.	Proposed-SIFT + LDA	68.21
3.	CNN + LDA	81.25
4.	Fine-tuning VGG16 + LDA	83.58
5.	Our approach	82.93

In order to have an intuitive feel for the scene recognition results, we created the following visualization table. In the table, each row represents a scene, and five different colors represent scenario name, test image, true positive with true predicted label, false positives with true label, and false positive with wrong predicted label. True positive with true predicted label represents that the image has an appropriate label and is assigned to the scene that belongs to it. False positives with true predicted label means that the image has a label that is consistent with it, but is incorrectly divided into another scene. For example, in the “museum” line, an image is classified as “museum” but it is actually “mall”. False positive with wrong predicted label describes an image belonging to a particular category that was incorrectly tagged with another scene. An example can be chosen, that is, in the “gym” row, an image belonging to “gym” is marked as “museum”. For the mispredicted images, the structure, layout, and other main features of the images are highly consistent with the scenes they are assigned to. That is, the topic with a higher probability in the image is similar to the topic of the scene it is misclassified, which leads to the emergence of the wrong predictions. Figure 11 shows the first four rows of the display table. Through the figure below, we can have an intuitive understanding of misclassified images and misidentified images.

scene	test image		true positive with true predicted label		false positive with true predicted label		false negative with wrong predicted label	
museum								
			museum	museum	mall	bowling	cloister	tallbuilding
gym								
			gym	gym	cloister	bocce	mall	museum
mall								
			mall	mall	insidecity	bowling	museum	tallbuilding
bowling								
			bowling	bowling	gym	badminton	mall	gym

**Figure 11.** Visual table in from combined datasets.

## 5. Conclusions

This paper proposes a method, named DF-LDA, for scene recognition that combines a convolutional neural network with a latent topic model. The purpose of this method is to combine the advantages of the two models. Among them, CNN can automatically learn better feature representations from the images and then use the topic models to capture the high-level scene structure of the images. The performance of this method has been further improved by using an enhanced clustering algorithm. Experiments conducted on two published benchmark datasets and a self-constructed dataset. The results demonstrate that the method can deliver higher-level topic semantics through the potential relevance of abstract visual features, which will be beneficial to improve the effects of scene recognition. This method overcomes the defects of the image content—such as image rotation, distortion, lighting, etc.—but if the image to be tested has problems such as occlusion and ghosting, it may not be able to play a good role. Because there are these issues, especially the key goals, it may not correctly infer useful subject information. In the future, we will consider further research on existing problems.

**Acknowledgments:** The work is supported by the National Nature Science Foundation of China (41571401).

**Author Contributions:** Jiangfan Feng and Amin Fu conceived the DF-LDA algorithms; Amin Fu performed the experiments. Both authors have read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
2. Lee, Y.-S.; Lo, R.; Chen, C.-Y.; Lin, P.-C.; Wang, J.-C. News topics categorization using latent Dirichlet allocation and sparse representation classifier. In Proceedings of the 2015 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taipei, Taiwan, 6–8 June 2015; pp. 136–137.
3. Ma, Y.; Jiang, Z.; Zhang, H.; Xie, F.; Zheng, Y.; Shi, H.; Zhao, Y. Breast histopathological image retrieval based on latent dirichlet allocation. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 1114–1123. [[CrossRef](#)] [[PubMed](#)]
4. Zhao, B.; Zhong, Y.; Xia, G.-S.; Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2108–2123. [[CrossRef](#)]
5. Tirunillai, S.; Tellis, G.J. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *J. Mark. Res.* **2014**, *51*, 463–479. [[CrossRef](#)]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS 2012), Proceedings of The Twenty-Sixth Annual Conference on Neural Information Processing Systems (NIPS), Stateline, NV, USA, 3–8 December 2012*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2012; pp. 1097–1105.
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
8. Quattoni, A.; Torralba, A. Recognizing indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 413–420.
9. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 30 July–1 August 1999; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp. 289–296.
10. Bosch, A.; Zisserman, A.; Muñoz, X. Scene classification via pLSA. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 517–530.
11. Fei-Fei, L.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–25 June 2005; pp. 524–531.
12. Niu, X.-X.; Suen, C.Y. A novel hybrid CNN-SVM classifier for recognizing handwritten digits. *Pattern Recognit.* **2012**, *45*, 1318–1325. [[CrossRef](#)]

13. Zhen, K.; Birla, M.; Crandall, D.; Zhang, B.; Qiu, J. Hybrid supervised-unsupervised image topic visualization with convolutional neural network and LDA. *arXiv*, **2017**, arXiv:preprint/1703.05243.
14. Wang, Z.; Wang, Y.; Wang, L.; Qiao, Y. Codebook enhancement of VLAD representation for visual recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 1258–1262.
15. Wang, Z.; Wang, L.; Wang, Y.; Zhang, B.; Qiao, Y. Weakly supervised PatchNets: Describing and aggregating local patches for scene recognition. *IEEE Trans. Image Process.* **2017**, *26*, 2028–2041. [[CrossRef](#)] [[PubMed](#)]
16. Wang, L.; Wang, Z.; Qiao, Y.; Van Gool, L. Transferring deep object and scene representations for event recognition in still images. *Int. J. Comput. Vis.* **2018**, *126*, 390–409. [[CrossRef](#)]
17. Li, Z.; Tian, W.; Li, Y.; Kuang, Z.; Liu, Y. A more effective method for image representation: Topic model based on latent dirichlet allocation. In Proceedings of the 2015 14th International Conference on Computer-Aided Design and Computer Graphics (CAD/Graphics), Xi'an, China, 26–28 August 2015; pp. 143–148.
18. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, USA, 23–28 June 2014; pp. 512–519.
19. Qiao, Y.; Wang, L.; Guo, S.; Wang, Z.; Huang, W.; Wang, Y. Good Practice on Deep Scene Classification: From Local Supervision to Knowledge Guided Disambiguation. Available online: <https://wangzheallen.github.io/papers/SceneGP.pdf> (accessed on 17 April 2018).
20. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2007; pp. 1027–1035.
21. Zhao, B.; Fei-Fei, L.; Xing, E.P. Image segmentation with topic random field. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 785–798.
22. Li, L.-J.; Li, F. What, where and who? Classifying events by scene and object recognition. In Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV 2007), Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
23. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [[CrossRef](#)]
24. Zang, M.; Wen, D.; Wang, K.; Liu, T.; Song, W. A novel topic feature for image scene classification. *Neurocomputing* **2015**, *148*, 467–476. [[CrossRef](#)]
25. Zang, M.; Wen, D.; Liu, T.; Zou, H.; Liu, C. A pooled object bank descriptor for image scene classification. *Expert Syst. Appl.* **2018**, *94*, 250–264. [[CrossRef](#)]
26. Jeon, J.; Kim, M. A spatial class LDA model for classification of sports scene images. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4649–4653.
27. Gao, S.; Tsang, I.W.-H.; Chia, L.-T.; Zhao, P. Local features are not lonely—Laplacian sparse coding for image classification. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3555–3561.

