Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549488

COMBINING HANDCRAFTED AND DEEP FEATURES FOR SCENE IMAGE CLASSIFICATION

Shrinivasa S.R. , Prabhakar C.J.
Department of Computer Science, Kuvempu University, Karnataka, India.
Corresponding Author : Shrinivasa S.R

ABSTRACT: *In the last decade, a plethora of handcrafted-based scene image classification techniques have been proposed in the literature. Some of them are based on structural analysis, while some others exploit mutual information or perceptual characteristics. Nowadays, deep learning-based methods are widely used in several domains due to its ability to well fit the target directly from the scene image. In this paper, an impact on the performance of combining handcrafted features and Deep features for scene image classification is addressed. In order to utilize the benefits of handcrafted and deep features, we extract and combine these features for scene image classification, which helps to improve the accuracy of the classification. We extract the deep features of the images based on the classical Res-Net model and handcrafted local features are extracted using Binary Robust Invariant Scalable Key points (BRISK) descriptor. By combining the two types of image features, we form a new type of image features, called hybrid features. Then, we employ Support Vector Machine (SVM) training model for scene image classification. We carried out the experiments on two public datasets such as CIFAR-10 and Corel-1000. The experimental results have shown that the proposed hybrid method has exhibited remarkable performance with high classification accuracy.*

I. INTRODUCTION

The broad range use of the Internet and the enormous use of audio-visual information in digital format for communications has created enormous amount of scene images belongs to both indoor and outdoor environments. The system for describing the content of hypermedia information in order to find and classify them is really important and difficult task. The scene image classification of scene images has been an active research topic. Scene image classification is to correctly label given scene images with predefined semantic categories. The scene image classification problem is not completely solved due to the performance hindered by a wide variety of challenges found in imbalanced and large scale datasets, such as illumination, perspective fluctuation, clutter, and occlusion [1]. The image descriptors have been increasing the accuracy and performance in scene image classification. The image descriptors describe primary characteristics such as shape, color, texture or motion of images, which are visual features of images. Most of the previously proposed methods for scene image classification systems have focused on using handcrafted image features, which are designed by expert knowledge of designers, such as SIFT, SURF, Gabor filter, local binary pattern (LBP), local ternary pattern (LTP), and histogram of oriented gradients (HOG). As a result, the extracted features reflect limited aspects of the problem, yielding a classification accuracy that is low and varies with the characteristics of scene images. Each of these descriptors has its own drawbacks such as higher dimension feature vectors and considering only certain features such

as texture and shape. To adapt to the various tasks being tackled by researchers, the number of features has increased.

The deep learning methods have been developed in the computer vision research community, which is proven to be suitable for automatically training a feature extractor that can be used to enhance the ability of handcrafted features. Deep Convolutional Neural Network (DCNN) is a method of learning image features with more discriminative and has been studied deeply and applied widely in the field of computer vision and pattern recognition. The researchers have proposed deep learning based methods using CNN-based features for image classification [2, 3, 4, 5, 6, 7]. The deep learning methods extract the image features using image filtering technique and a neural network to classify the extracted image features into several desired classes. Using considerable training data, the deep learning methods have proven to outperform conventional methods in scene classification methods. The literature survey reveals that the deep learning framework was applied for scene image classification problem as an image feature extractor. They used several CNN models which were trained to extract image features. They additionally used several kinds of handcrafted image feature extraction methods such as LBP, SIFT, SURF, HOG and local ternary pattern (LTP), to extract the image features besides the deep features for scene image classification. As a result of this study, they proved that the handcrafted and deep image features can extract different information from input images. Based on this result, they showed that the combination of handcrafted and deep features is sufficient for enhancing the classification accuracy. However, the methods proposed in these studies use multiple CNN models and methods for handcrafted image feature extraction.

To overcome the limitations of previously proposed scene classification methods, we propose a method that uses a combination of deep and handcrafted features extracted from the images. Our proposed method uses the pre-trained ResNet convolutional neural network (ResNet CNN) to extract deep image features and the Binary Robust Invariant Scalable Key points (BRISK) descriptor to extract detailed features from scene images. By combining the two types of image features, we form a new type of image features, called hybrid features, which has stronger discrimination ability than single image features. Finally, we use the support vector machine (SVM) method to classify the image features into one of the categories. Our experimental results indicate that our proposed method outperforms previous scene classification methods by yielding the smallest error rates on the same image databases.

II. Related Work

Scene image classification is an ongoing field of study in computer vision communities due to its wide range of applications and challenges posed by this problem. The traditional methods use hand-crafted features for Scene image classification based on various descriptors and its variants. However, the performances of these methods are limited by low-capacity features. The development of deep learning technology has led to advancements in innovation and efficiency of various methods. Comparing with traditional methods, deep learning methods have shown highest accuracy. In this section, we discuss the progress in the field of scene image classification using handcrafted and deep features.

1.1 Handcrafted features

In the past decades, many methods have been presented for scene classification. These methods mainly focus on designing various human-engineering either local or global features, such as color, texture, and shape information or their combination, which are the primary characteristics of a scene image. Some of these methods use local descriptors. For example, the speeded up robust feature(SURF) and scale invariant feature transform (SIFT) [6] for describing local structural variations in scene image. In addition, distributions exploitation on certain spatial cues such as color histogram [7], texture information [8] has also been well surveyed. In [9], local structural texture similarity descriptor was applied to image blocks to represent structural texture for image classification. In [10], semantic classification of scenes based on Gabor and Gist descriptors [11] was evaluated individually. In order to depict the complex scene, the combinations of complementary features are often preferred to achieve improved results. In [12], different kinds of feature descriptors, i.e., Gaussian wavelet features, gray level co-occurrence matrix, Gabor filters and shape features, were combined to form a multiple-feature representation for indexing scene images with different spatial resolutions, and better performance was reported.

Recently, local binary pattern (LBP) [13] and completed LBP (CLBP) [14] are also presented. Afterwards, multiscale completed LBP (MS-CLBP) [15] and extended multistructure LBPs (EMSLBP) [16] were adopted for image scene classification and competitive results were reported. However, this kind of methods may not able to produce discriminative representation, especially when salient structures in complex background scene images often dominate the image classification. For instance, Yang et al. [8] extracted SURF and Gabor texture features for classifying scene images and demonstrated SIFT performs better. However, one limitation of the methods that use the local descriptors is a lack of the global distributions of spatial cues. In order to depict the spatial arrangements of images, Santos et al. [9] evaluated various global color descriptors and texture descriptors, for example, color histogram for scene image classification. To further improve the classification performance, Luo et al. [14] combined different types of feature descriptors, including local and global descriptors, to form a multi-feature representation for describing scene images. However, in practical applications, the performance is largely limited by the hand-crafted descriptors, as these make it difficult to capture the rich semantic information contained in scene images.

Because of the limited discrimination of hand-crafted features, these methods mainly attempt to develop a set of basic functions used for feature encoding. One of the most popular mid-level approaches is the bag-of-visual-words (BoVW) model [20]. However, the BoVW-based models may not fully exploit spatial information and poor representation capability for scene classification. Although these methods have not made achievements in scene image classification, they all demand prior knowledge in handcrafted feature extraction. Further, these approaches based on hand-crafted features may not well represent semantic information. A major shortcoming of these hand-crafted features is that they demand complex engineering skills that rely on expert experience. However, we utilized deep learning methods among hand-crafted feature techniques and representations, as they have shown their notable performance for the image classification task.

1.2 Deep features

The rapid development of deep learning technologies accelerates the progress in scene classification using Convolutional Neural Networks (CNN) which has been widely applied to image classification [16], object detection [17], semantic segmentation [18]. Many scene image classification methods employ pre-trained networks on the ImageNet dataset [17] as feature extractors such as VGGNet [19], AlexNet [20], Google Net [21]. And not only that, there are many novel networks are designed for scene image classification [22] [23]. Wang et al. [24] employed the rich hierarchical features of a CNN to form a discriminative image representation for scene image classification which incorporates low-level and high-level features simultaneously. Liu et al. [25] introduced a Siamese CNN model that combines verification and identification models to boost the performance. To allow the input images to be of arbitrary sizes, Xie et al. [26] proposed a scale-free CNN topreserve key information in high complex background information images.

A multi-scale CNN [27] is proposed to merge the feature maps of different layers based on feature maps selection algorithm and region covariance descriptor. Wang et al. [28] utilized attention mechanism to adaptively select a series of critical parts of scene images, and then to generate powerful features. Bi et al. [29] proposed a multiple-instance learning (MIL) framework to highlight the local semantics relevant to the scene image label. Scene classification is a challenging task due to the inter and intra class diversity, and complex spatial distributions of scene images [30], which result in large intra-class variance and small inter-class variance. Although these existing CNN-based methods have achieved great performance to some extent, some complex scene classes are still easily misclassified since only visual information is utilized. The most of the previous methods [22][23] only learn the global features representation of images, which may neglect the local details. Even though there are several works [24][29] attempt to focus on the critical local image patches and discard the useless information, they still only utilize the visual information.

III. METHODOLOGY

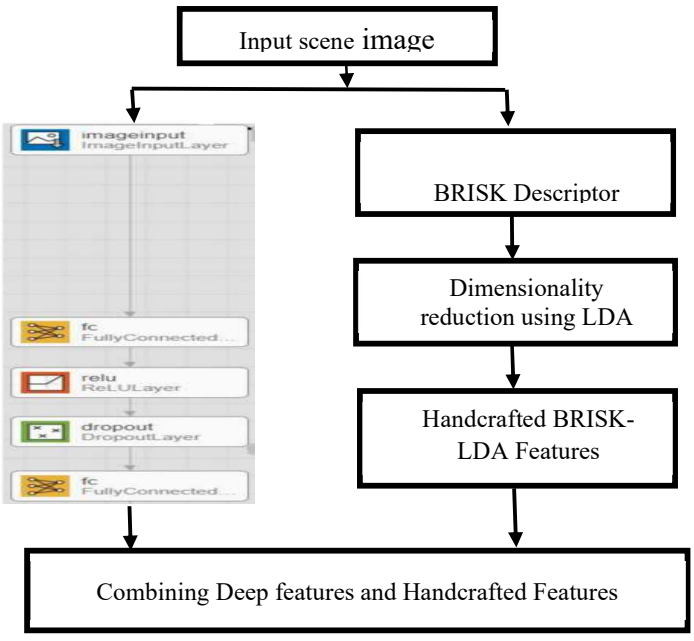

Fig.1 Flow diagram of the proposed framework for scene image classification

In this section, we present proposed method for scene image classification by combining handcrafted and deep features. The flow of the proposed method is shown in the Figure 1. The training images are resized to 224× 224 pixels. Then, each image is fed into pre-trained deep feature extractor (i.e. ResNet-50 CNN) where it extracts deep features from an image. Similarly, handcrafted features such as BRISK descriptors are extracted from training image. In order to reduce the dimension of the BRISK descriptor, we employ Linear Dimensionality Analysis (LDA), where, 64 of 128 features are selected and a handcrafted BRISK- LDA feature vectors with a dimension of 1 × 64 are created. By combining the two types of image features, we form a new type of image features, called hybrid features, which has stronger discrimination ability than single image features. Finally, we use the support vector machine (SVM) method to classify the image features into one of the categories.

3.1 BRISK- Handcrafted Features Descriptor

The BRISK algorithm [31] is a feature point detection and description algorithm with scale invariance and rotation invariance. It constructs the feature descriptor of the local image through the gray scale relationship of random point pairs in the neighborhood of the local image, and obtains the binary feature descriptor. Compared with the traditional algorithm, the matching speed of BRISK is faster and the storage memory is lower. The BRISK algorithm includes two main modules: keypoints detection, and keypoints description. First, the scale space pyramid is constructed, and the stable extreme points of sub-pixel precision in continuous scale space are extracted by the Adaptive corner detection operator. The key concept of the BRISK descriptor makes use of a pattern used for sampling the neighborhood of the keypoint. The pattern, illustrated in Figure 2, defines N locations equally spaced on circles concentric with the keypoint. In order to avoid aliasing effects when sampling the image intensity of a point in the pattern, Gaussian smoothing is done. Then, the binary feature descriptor of the local image is established by using the gray scale relationship of the random sample point pairs in the local image neighborhood.

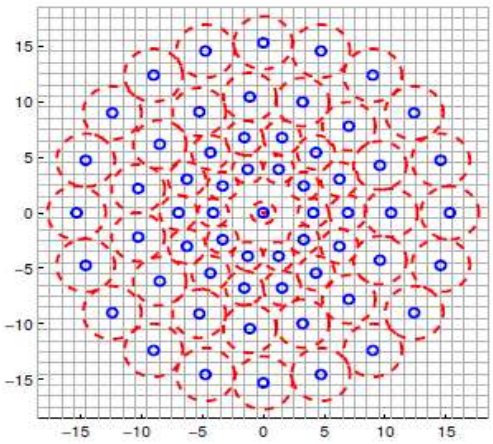

Fig.2. BRISK descriptor – BRISK sampling pattern

The feature extraction is a method utilized to recognize or extract key parameters from input images as initial information to obtain the novel data. In this work, key pointer detection method is applied to detect the features and also locate the extracted features via key points. Especially, BRISK algorithm is utilized which is an attribute point recognition as well as depiction approach with scalable invariance along with turning round. Moreover, this detector contest attributes among two images via tuning the parameters up to 200 values. The group of key points comprises of points cultures image positions linked with floating point scaled principles. The BRISK descriptor is collected as two-fold string through adding the outcomes of effortless intensity of image comparison trials. This intensity comparison helps to enhance the image descriptiveness.

Fig. 2 depicts the BRISK descriptor used for feature extraction. BRISK descriptor is applied in [20] to identify key points, then description and finally undergo matching process. The current descriptor which is used for feature extraction consists of coaxial rings. We should consider little scrap of the coaxial rings and yet to apply Gaussian method for smoothening the brain images while we are taking every point in the circle. The red color mentioned in the circle represented how long the divergence of filter took place in every point. It captures the group of undersized couples, spins the couples by point of reference evaluated and then creates assessment in the form of Eq. (1) and Eq. (2).

$$g(\mathrm{p}_i, \mathrm{p}_j) = \mathrm{p}_j - p_i \cdot \frac{I(\mathrm{p}_j, \sigma_j) - I(\mathrm{p}_i, \sigma_i)}{\left\| p_j - p_i \right\|^2} \tag{1}$$

$$g = \begin{pmatrix} g_x \\ g_y \end{pmatrix} = \frac{1}{L} \cdot \sum_{(\mathrm{p}_i, \mathrm{p}_j) \in L} g(\mathrm{p}_i, \mathrm{p}_j). \tag{2}$$

3.2 ResNet-50 - Deep features

ResNet stands for Residual Network and is a specific type of convolutional neural network (CNN) introduced in the 2015 by [32]. ResNet-50 is a 50-layer convolutional neural network (48 convolutional layers, one Max Pool layer, and one average pool layer). Over the year's deep convolutional neural networks have made a series of breakthroughs in the field of image recognition and classification. Going deeper to solve more complex tasks and to improve classification or recognition accuracy has become a trend. But, training deeper neural networks has been difficult due to problems such as vanishing gradient problem [27] and degradation problem [28]. Residual learning tries to solve both these problems. In neural networks every layer learns low or high level features while being trained for the task at hand. In residual learning instead of trying to learn features, model tries to learn some residual. As we can see in Fig. 8, the input 'x' is being added as a residue to the output of the weight layers and the activation is carried out. Relu activations are being used in the ResNet model. ResNet50 is a 50-layer Residual network and has other variants such as ResNet101 and ResNet152. Using ResNet as a pre trained model for scene image classification has brought good results [29].

ResNet-50 Architecture

Deep Residual Network is almost similar to the networks which have convolution, pooling, activation and fully-connected layers stacked one over the other. ResNet have something called Residual blocks. Many Residual blocks are stacked together to form a ResNet. The skipped connections which are the major part of ResNet. The idea is to connect the input of a layer directly to the output of a layer after skipping a few connections. We can see here, x is the input to the layer which we are directly using to connect to a layer after skipping the identity connections and if we think the output from identity connection to be F(x). Then we can say the output will be

F(x) + x.

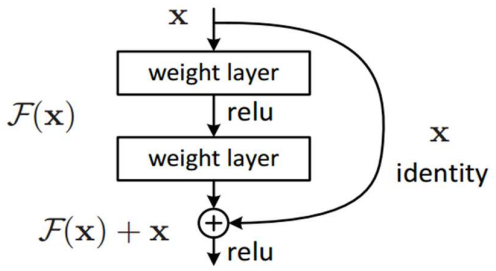

Fig.3. Residual Learning – Skip connection

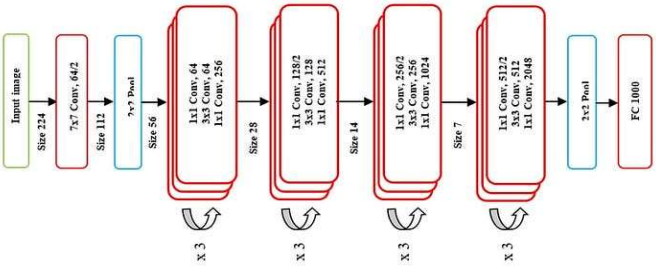

Fig.4. ResNet-50 Architecture

The 50-layer ResNet architecture includes the following elements, as shown below:

- 7×7 kernel convolution alongside 64 other kernels with a 2-sized stride.
- A max pooling layer with a 2-sized stride.
- 9 more layers—3×3, 64 kernel convolution, another with 1×1, 64 kernels, and a third with 1×1, 256 kernels. These 3 layers are repeated 3 times.
- 12 more layers with 1×1, 128 kernels, 3×3, 128 kernels, and 1×1, 512 kernels, iterated 4 times.
- 18 more layers with 1×1, 256 cores, and 2 cores 3×3, 256 and 1×1, 1024, iterated 6 times.
- 9 more layers with 1×1, 512 cores, 3×3, 512 cores, and 1×1, 2048 cores iterated 3 times.
- Average pooling, followed by a fully connected layer with 1000 nodes, using the softmax activation function.

- In transfer learning firstly a base network is trained on a base dataset and the features learned from the first task are repurposed or transferred to a second network to train on a second dataset and task. This process will work if the features are suitable for both base and target tasks, instead of only base task [18]. Deploying pre-trained models on similar data have shown good results in image classification related tasks [19-20]. Few organizations have created models such as Oxford VGG Model [21], Google Inception Model [22] and Microsoft ResNet Model [23] which take weeks to train on modern hardware. These models can be downloaded and integrate with new models which take image as input to bring better results.

(1) Input layer. The training images are resized to 224× 224 pixels. Then, each image is fed into pre-trained deep feature extractor (i.e. ResNet CNN) where it extracts deep features from an image.

(2) Convolutional layer (C1). C1 is the first convolutional layer of the feature extraction, and we obtain 96 feature maps with the size of 55×55. In fact, it is obtained by utilizing the convolutional kernel of size 11×11. Because the size of the receptive field of the neuron is determined by the size of the convolutional kernel, the ideal size of the convolutional kernel is to extract the effective local features in the range of convolutional kernel with representation ability. Therefore, the proper setting of the convolutional kernel is very important to extract the effective image features and improve the performance of the convolutional neural network. C1 filters the 224×224 input image with 96 convolutional kernels of size 11×11 with a stride of 4 pixels for sampling frequency, namely the convolutional kernel is spread over every unit of size 11×11. Ultimately, we get 96 feature maps with the size of (224/4-1) × (224/4-1) =55×55.

(3) Max-pooling layer (P1). P1 is the first max-pooling layer, which has 96 feature maps of size 27×27. The pooling process is to select the maximum in each of the pooling regions as the value of the area after pooling. In this layer, we choose a max-pooling layer over a 3×3 region in order to control the speed of dimensionality reduction, because the decline in dimension is decreased exponentially, the speed is falling faster means the image features are more rough, and many image details are lost subsequently. Since the pooling layer has a region of size 3×3, while the stride of size 2, so we obtain the overlapping pooling. This scheme reduces the top-1 and top-5 error rates by 0.4% and 0.3%, respectively, as compared with the no-overlapping scheme, which produces output of equivalent dimensions.

(4) Convolutional layer (C2). The second feature extraction layer C2, which has a lot of similarities with the C1, while there is a certain gap. C2 takes as input the output of the first convolutional layer and filters it with 256 convolutional kernels of size 5×5 with a stride of 1pixel, and it is feasible to obtain 256 feature maps with the size of 27×27. In C1 layer, the receptive field of each neuron is equivalent to the raw image of size 33×33. Now C2 layer use a kernel of 5×5 to convolution, so that the receptive fields of neurons further enhance, which is equivalent to the original image of size 165×165. Each feature maps in C2 is not directly obtained by convoluting in P1 layer, but by combining several or all of the feature maps in P1 as input for convoluting again. The reason for this one is that the sparsely connected mechanism keeps the number of connections in a reasonable range. The other is that the asymmetry of network enables the different combinations can extract various features.

(5) The remaining convolutional layers and pooling layers. These layers have the same working principle with the first two layers, but the size and number of the feature maps have changed. However, the size of the C5 layer after the convolution is still 13×13. In this process, the max-pooling layers follow the second (C2) and fifth (C5) convolutional layers with the kernels of size 3×3. With the continuous increase of the depth of the convolution, the extracted features are more abstract, with more discriminative and expressive power. The experimental results demonstrate that the classification accuracy is only about 40% when only take the first two layers of the convolutional neural network, which proves that depth will have a great effect on the performance of convolutional neural network, and lacks of depth will reduce the abilities of extracting features in the convolutional neural network.

(6) Fully-connected layer (Fc). In the whole process of learning, the remaining three are the fully-connected layers, which make the learning features of two channels cross-mixing to obtain a 4096-dimensional feature vector. We use Dropout technique in the first two fully-connected layers, setting to zero the input of the first two fully-connected layers with the probability 1/2, which do not contribute to the forward pass and do not participate in backward. The hidden layer of learning in this way cannot rely on the presence of particular other features of the previous layer, which makes the learning features can lead to stronger robustness, select more adaptive parameters, and significantly improve the generalization capabilities of the system.

Table (1): The optimal parameters for ResNet 50

Parameter	**Value**
Epochs	100
Validation step	1
Optimizer	SGDM (Stochastic Gradient Descent with Momentum)
Learning rate	Piecewise scheduler
Decay	Default
Momentum	Default

IV. EXPERIMENTAL RESULTS

In this section, we present experimental results obtained for our proposed method using popular datasets such as CIFAR-100 [33] and Corel-1000 [34] are presented. All the experiments have been performed using Intel Core-i7 CPU with 2.7 GHz, 8GB RAM. The training and testing ratio of 70:30 is used for all experiments. We used metrics for evaluation of classification performance is the classification accuracy (A), defined as total instances (images) correctly classified and fractionated by total number of instances (images) within the dataset under consideration. It is mathematically expressed as

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (3)$$

To analyze the effectiveness of the implemented technique, diverse image classification benchmarks which are widely used in literature have been utilized. In below summarizes details regarding the total number of classes, images per class, number of images per class and total number of images in the benchmark, image spatial resolution, and dimensions. We evaluated our proposed method using two benchmark datasets, CIFAR-10 [33] and Corel-1000 [34].

Experiments on CIFAR-10 dataset

The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. Figure 5 shows some example images. The dataset is divided into 50000 training images and 10000 test images. The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

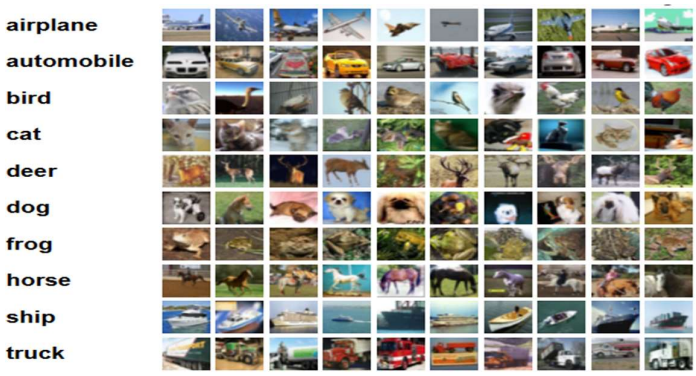

Fig.5. Sample images from CIFAR-10 dataset

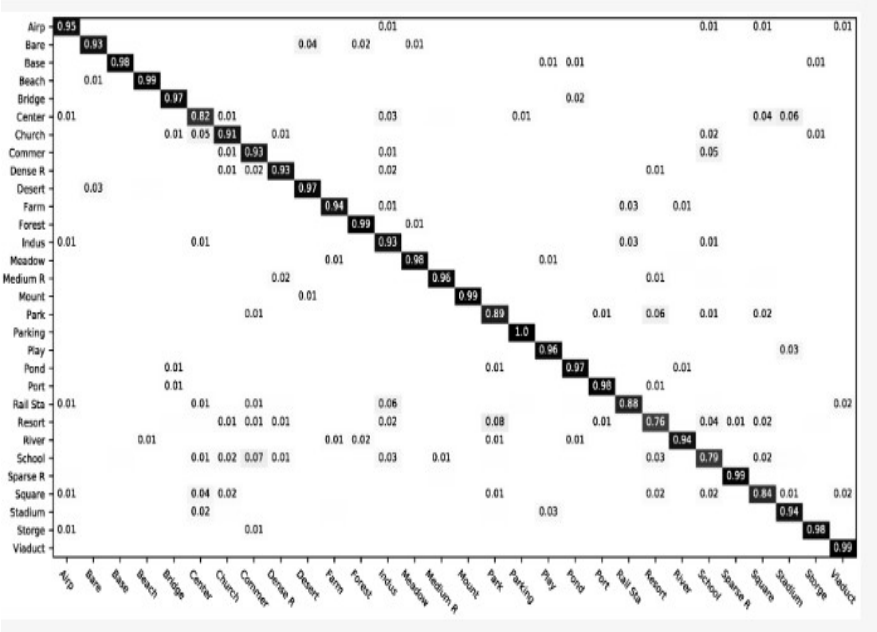

Fig.6: Confusion matrix for proposed method on CIFAR-10 dataset.

The Fig. 6 shows the confusion matrix generated by the proposed method. From the confusion matrix, we can see the 10 categories achieved the classification accuracy of 98%. The dataset are very challenging like inter-class dissimilarity, rich semantic and complex background could also be accurately classified. However, the major confusions were between school and commercial, resort and park. As illustrated in Fig.6, school and commercial have the same image distribution, for example, clutter structures; resort and park have the analogous objects and image texture, for example, green belts and buildings. Thus, these classes were easily confused. Even so, our method achieved a substantial improvement for the difficult scene types show in accuracies (0.49, 0.6, 0.63, 0.65) of the same classes from the confusion matrix. This result is possibly explained by the fact that the combining of hand crafted features and deep features give the ability to learn discriminative features. Particularly, for the scenes that are rich in obvious objects, such as airport, industry, and dense residential, our method can achieve higher accuracies and comparable performance. Thus, the combining of the hand crafted features and deep features can achieve accurate scene reasoning.

Experiments on Corel-1000 dataset

The Corel-1000 dataset consists 10,800 images belongs to 80 classes; each class includes more than 100 images. Figure.7 shows some example images. The dataset is divided into 9000 training images and 1800 test images. The experimental setup is same for as mentioned in CIFAR-10 dataset.

Fig.6. Sample images from Corel-1000 dataset

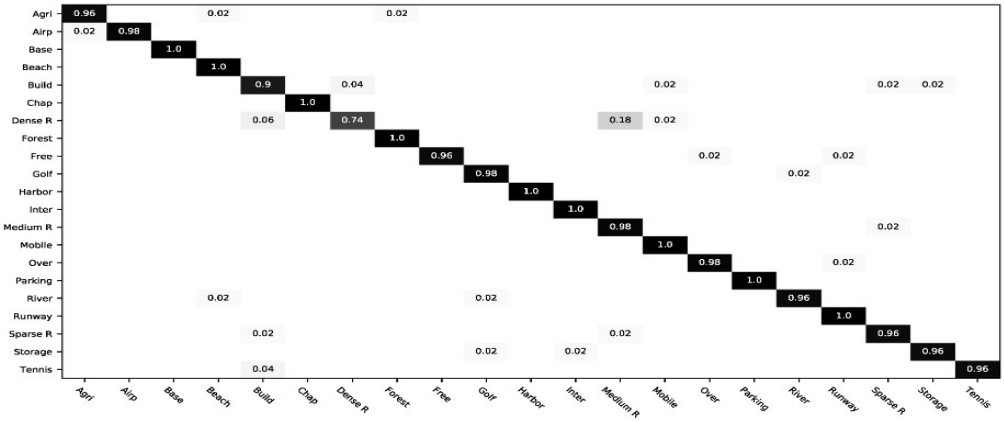

Fig.8: Confusion matrix of our method on Corel-1000 dataset.

The Fig. 8 shows the confusion matrix generated by the proposed method. From the confusion matrix, we can see that the classification accuracy of 96%. The dataset are very challenging like inter-class dissimilarity, rich semantic and complex background could also be accurately classified. The results demonstrate that combining of hand crafted features and deep features give the ability to learn discriminative features.

Comparative Study

We used Precision, Recall and classification accuracy metrics in order to evaluate the proposed method using two public datasets. We compared the proposed method with Handcrafted features based techniques such as BRISK [35] and SURF and DBC [36]. Similarly, the performance of the proposed method is compared with the deep features extracted using AlexNet [37] and ResNet-50 [38] model. The classification accuracy, precision, and recall on the above two public datasets, obtained by the proposed method and the existing state-of-the-arts techniques including the approaches based on deep models are shown in Table 2. These results prove that the proposed framework significantly improves the performance of scene image classification even in the presence of the blurred, low illumination, orientation and complex background in real scenes. From the results demonstrated in Table 2 on the CIFAR-10 and Corel-1000 datasets, it can be seen that the proposed framework outperforms existing methods for scene image classification.

Table (2): Comparison of the scene image classification results of the proposed method with the state-of-the art methods on the CIFAR-1 and Corel-1000 datasets.

Methods	CIFAR-10			Corel-1000		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
BRISK-LDA [35]	0.761	0.612	76.12	0.561	0.684	78.56
SURF and DBC [36]	0.811	0.121	91.21	0.839	0.841	93.22
AlexNet [37]	0.912	0.835	95.56	0.756	0.891	96.36
ResNet-50 [38]	0.812	0.941	97.35	0.893	0.951	94.51
Proposed Method	0.995	0.991	99.12	0.914	0.937	99.70

The improved performance of the proposed method is due to the following reasons:

1. The feature selection using LDA algorithm has been utilized in order to decrease training time, computing volume, and to select discriminative features.
2. BRISK descriptor has the advantage of being performed on local cells that are invariant to geometric and brightness conversions except for the orientation of the object.
3. The ResNet-50 CNN detects the important and high level features automatically without any human supervision.
4. Instead of considering handcrafted features alone and deep features alone, we combined the handcrafted and deep features which were extracted using BRISK and ResNet-50.

5. Conclusion

In this paper, an efficient method for scene image classification is proposed based on hybrid features through combination of Handcrafted features and deep features. We extracted handcrafted features using BRISK descriptor and deep features are extracted using ResNet-50 model. The experiments conducted on two datasets such as CIFAR-10 and Corel-1000 datasets demonstrates that the proposed method outperforms existing methods for scene classification. It is obvious that the proposed method is an efficient way for scene image classification, hence, it is suitable for scene image classification. One of the merits of proposed method is sensitivity to intra-class and inter-class variety when it is used for classification. In other words, related images with more similarity are classified early. Another advantage is high performance on imbalanced databases. Plus, the hybrid features through combination of handcraft features deep features achieve superior accuracy using SVM classifier.

REFRENCES

[1]. Banerji, S., Sinha, A., & Liu, C. (2013). New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. Neurocomputing, 117, 173-185.

[2]. Giveki, D., Soltanshahi, M. A., & Montazer, G. A. (2017). A new image feature descriptor for content based image retrieval using scale invariant feature transform and local derivative pattern. Optik, 131, 242-254.

[3]. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN Architecture for ´ Weakly Supervised Place Recognition. In: Proc. CVPR (2016).

[4]. Lee, Y., Lim, S., & Kwak, I. Y. (2021). Cnn-based acoustic scene classification system. Electronics, 10(4), 371.

[5]. Cheng, G., Xie, X., Han, J., Guo, L., & Xia, G. S. (2020). Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 3735-3756.

[6]. Zeng, D., Liao, M., Tavakolian, M., Guo, Y., Zhou, B., Hu, D., ... & Liu, L. (2021). Deep learning for scene classification: A survey. arXiv preprint arXiv:2101.10531.

[7]. Liu, S., Tian, G., & Xu, Y. (2019). A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. Neurocomputing, 338, 191-206.]

[8]. N. Dalal and B. Triggs, “Histograms of oriented gradients for human de- tection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., San Diego, CA, USA, Jun. 2005, pp. 886-893.

[9]. J. Ren, X. Jiang and J. Yuan, “Learning LBP structure by maximizing the conditional mutual information,” Pattern Recognit, vol. 48, pp. 3180- 3190, Feb. 2015.

[10]. D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, pp. 91-110, Nov. 2004.

[11]. L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., San Diego, CA, USA, Jun. 2005, pp. 524-531.

[12]. S. Lazebnik, C. Schmid and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., New York, NY, USA, Jun. 2006, pp. 2169-2178.

[13]. H. Jgou, F. Perronnin, M. Douze, J. Snchez, P. Prez and C. Schmid, “Aggregating Local Image Descriptors into Compact Codes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, pp. 1704-1716, Sept. 2011.

[14]. B. Zhao, Y. Zhong, L. Zhang, and B. Huang, “The fisher kernel coding framework for high spatial resolution scene classification,” Remote Sens., vol. 8, no. 2, p. 157, Feb 2016.

[15]. Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., San Jose, CA, USA, Nov. 2010, pp. 270-279.

[16]. Q. Zhu, Y. Zhong, B. Zhao, G.S. Xia, and L. Zhang, "Bag-of-visualwords scene classifier with local and global features for high spatial resolution remote sensing imagery," IEEE Geosci. Remote Sens. Lett., vol. 13, no. 6, pp. 747-751, Jun. 2016.

[17]. K.M. He, X.Y. Zhang, S.Q. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, Jun. 2016, pp. 770-778.

[18]. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Ohio, CO, USA, Jun. 2014, pp. 580-587.

[19]. J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, Jun. 2015, pp. 3431-3440.

[20]. J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, Jun. 2009, pp. 248-255.

[21]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations., SanDiego, CA, USA, May. 2015.

[22]. Krizhevsky, I. Sutskever, and G.E.Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural. Inf. Process. Syst., Lake Tahoe, NY, USA, Dec. 2012, pp. 1097-1105.

[23]. Szegedy, W. Liu and Y. Jia, "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, Jun. 2015; pp. 1-9.

[24]. F. Luus, B. Salmon, F. Van Den Bergh, and B. Maharaj, "Multiview deep learning for land-use classification," IEEE Geosci. Remote Sens. Lett., vol. 12, no. 12, pp. 2448-2452, Dec. 2015.

[25]. Y. Liu, Y. Zhong and Q. Qin, "Scene Classification Based on Multiscale Convolutional Neural Network," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9, pp. 7109-7121, Sept. 2018.

[26]. G. L. Wang, B. Fan, S. M. Xiang, and C. Pan, "Aggregating rich hierarchical features for scene classification in remote sensing imagery," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 9, pp. 4104-4115, Sep. 2017.

[27]. J. Xie, N. He, L. Fang and A. Plaza, "Scale-Free Convolutional Neural Network for Remote Sensing Scene Classification," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 9, pp. 6916-6928, Sept. 2019.Traore,

[28]. Y. Liu, Y. Zhong and Q. Qin, "Scene Classification Based on Multiscale Convolutional Neural Network," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9, pp. 7109-7121, Sept. 2018.

[29]. Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu and G. Xia, "A MultipleInstance Densely-Connected ConvNet for Aerial Scene Classification," IEEE Trans. Image Process., vol. 29, pp. 4911-4926, 2020.

[30]. Kamsu-Foguem, B., & Tangara, F. (2018). Deep convolution neural network for image recognition. Ecological informatics, 48, 257-268.

[31]. Leutenegger S., Chli M., Siegwart R.Y. BRISK: Binary Robust invariant scalable keypoints; Proceedings of the 2011 International Conference on Computer Vision; Barcelona, Spain. 6–13 November 2011; pp. 2548–255.

[32]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).5.

[33]. Krizhevsky, A., & Hinton, G. (2010). Convolutional deep belief networks on cifar-10. Unpublished manuscript, 40(7), 1-9.

[34]. Wang, Jing & Wang, Lidong & Liu, Xiaodong & Ren, Yan & Yuan, Ye. (2018). Color-Based Image Retrieval Using Proximity Space Theory. Algorithms. 11. 115. 10.3390/a11080115.

[35]. N. Rasiwasia and N. Vasconcelos, "Latent Dirichlet Allocation Models for Image Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2665-2679, Nov. 2013, doi: 10.1109/TPAMI.2013.69.

[36]. Shrinivasa, S. R., & Prabhakar, C. J. (2022). Scene image classification based on visual words concatenation of local and global features. Multimedia Tools and Applications, 1-20.

[37]. Sun, J., Cai, X., Sun, F., & Zhang, J. (2016, August). Scene image classification method based on Alex-Net model. In 2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS) (pp. 363-367). IEEE.

[38]. Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., ... & Dar, S. H. (2021). Satellite and scene image classification based on transfer learning and fine tuning of ResNet50. Mathematical Problems in Engineering, 2021, 1-18.