Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Scene Recognition by Joint Learning of DNN from Bag of Visual Words and Convolutional DCT Features

Abdul Rehman, Summra Saleem, Usman Ghani Khan, Saira Jabeen & M. Omair Shafiq

To cite this article: Abdul Rehman, Summra Saleem, Usman Ghani Khan, Saira Jabeen & M. Omair Shafiq (2021) Scene Recognition by Joint Learning of DNN from Bag of Visual Words and Convolutional DCT Features, Applied Artificial Intelligence, 35:9, 623-641, DOI: 10.1080/08839514.2021.1881296

To link to this article: https://doi.org/10.1080/08839514.2021.1881296

Published online: 25 May 2021.

APPLIED ARTIFICIAL INTELLIGENCE
2021, VOL. 35, NO. 9, 623–641
https://doi.org/10.1080/08839514.2021.1881296

Taylor & Francis
Taylor & Francis Group

Scene Recognition by Joint Learning of DNN from Bag of Visual Words and Convolutional DCT Features

Abdul Rehman[a], Summra Saleem[a], Usman Ghani Khan[a], Saira Jabeen[a], and M. Omair Shafiq[b]

[a]Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan; [b]School of Information Technology, Carleton University, Ottawa, Ontario, Canada

ABSTRACT

Scene recognition is used in many computer vision and related applications, including information retrieval, robotics, real-time monitoring, and event-classification. Due to the complex nature of the task of scene recognition, it has been greatly improved by deep learning architectures that can be trained by utilizing large and comprehensive datasets. This paper presents a scene classification method in which local and global features are used and are concatenated with the DCT-Convolutional features of AlexNet. The features are fed into AlexNet's fully connected layers for classification. The local and global features are made efficient by selecting the correct size of Bag of Visual Words (BOVW) and feature selection techniques, which are evaluated in the experimentation section. We used AlexNet with the modification of adding additional dense fully connected layers and compared its result with the model previously trained on the Places365 dataset. Our model is also compared with other scene recognition methods, and it clearly outperforms in terms of accuracy.

ARTICLE HISTORY
Received 20 May 2021
Accepted 24 May 2021

Introduction

In the recent years, the quick developments in innovation of technology has brought an exponential growth of media information due to which the data in the multimedia category (i.e. images, videos, etc.) is increasing exponentially. Scene classification has become significant by rapid increase in multimedia signal processing and applications that encounter storage and retrieval from large image databases based on image attributes. Nowadays, scene classification has vast applications in the field of computer vision and image processing. For example, the indoor-outdoor classification of an image can be used in image orientation detection (Bianco et al. 2008), Query by Image Content (QBIC), Content-Based Visual Information Retrieval (CBVIR), Robotics

CONTACT Usman Ghani Khan usman.ghani@kics.edu.pk Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan; M. M. Omair Shafiq omair.shafiq@carleton.ca School of Information Technology, Carleton University, Ottawa, Ontario, Canada
This article has been republished with minor changes. These changes do not impact the academic content of the article.

(Kim, Park, and Kim 2010), and event classification. Due to increasing concerns of security, this problem of scene classification demands a viable solution to save the community from the devastating effects of damage because of poor security. But these applications require further sub-classification of images that is the retrieval of images based on contents.

Content-based search of images rely on the contents of the image instead of metadata. The term content in this context is used to refer to color, structure, shape, texture, or any other visual information that can be extracted or derived from the image. Characterization of scenes, for example classification of mountains, forests, and workplaces is not a simple task. Undertaking attributable to this changeability, vagueness, and extensive variety of light and scale conditions that may apply. Semantic objects alone cannot assist much in the classification of a scene, e.g. a swimming pool can be indoor or outdoor.

But as far as the objects are concerned, they share the same semantic objects which is this case is a pool. However, they ought to be arranged uniquely in contrast to the part of indoor/outdoor scene characterization. Also, due to overlapping feature sets of indoor and outdoor images, achieving higher accuracy is a bit challenging task. Owing to various uses, convolution neutral networks are nowadays applied in many applications. On computer vision tasks like classification/detection of objects, characters, they have been proved quite efficient. The drawback of these models is that millions of parameter are involved depending upon the complexity of the network. The problem with complex networks is the requirement of large storage and a large amount of training epochs for efficient working, as with a deeper network, the number of layers increases and the resultant number of nodes per layer.

Geographical scene classification can be used as an efficient tool in the analysis of spatial data, environmental management, indoor and outdoor mapping. For refining the image and enhancing classification accuracies, spatial features have been recognized to be valuable. Practically, the method used commonly for combining local spatial features, the arrangement of appearance, and spatial images has been named as "Bag of Visual words" (BoVW). This basic prerequisite of this model is to identify the spatial features per pixel using a small number of training samples. Although the method has its own advantages in small data sets, but for larger ones, its performance is inconsistent. Hence, with higher geographical imagery, it is difficult to obtain higher accuracies. The solution to this lies in combining both CNN- and BoVW-based images in the classification of the geographical scene. Feature learning task requirement can be achieved by CNN which has biologically inspired trained architecture with the ability to learn multilevel hierarchies of features.

The proposed research work enhances the classification capability of network by utilizing and synergizing the existing state of the art techniques and incorporating DCT and BoVW features in the network structure. This study is

based on combining a CNN-based feature extractor employed on the DCT of the image with the BoVW-based scene classification. For adaptive and practical tracking of BoVW spatial feature Extractor, CNN serves the purpose. Training CNN with DCT features of data provides the required spatial features. The benefit of using a combined CNN and BoVW based feature extractor is to be able to acquire the core properties of original data. Also in experiment, it is evident that this method is consistent with various transforms generating accurate information with greater efficiency of scene classification.

Literature Survey

Researchers have been making contributions towards scene understanding for many decades. Attempts have been made from low-level feature extraction to global features and automatic feature extraction, i.e., through deep learning techniques.

Low-level Features Based Techniques

Raja, Roomi, and Dharmalakshmi (2014) proposed statistical features computed by using HSV (hue, saturation, value) color model as a color feature, DCT coefficient values as texture features, and entropy values of UV (chrominance) followed by the classification through KNN (k-nearest neighbor) classifier. LST (luminance, chrominance) as color and wavelet decomposition for texture feature used by Serrano, Savakis, and Luo (2002). Szummer and Picard (1998) used Ohta, Kanade, and Sakai (1980) color space to obtain color feature, multi-resolution simultaneous autoregressive model (MSAR) parameters, and coefficients of DCT to obtain texture features. Later on, people moved towards global features such as SIFT features followed by SPM (spatial pyramid matching) which achieved significant performance in scene classification (Lazebnik, Schmid, and Ponce 2006).

While Feng, Liu, and Wu (2017) argued that rapid recognition of scene is not only based on the top of object items but it may be examined in parallel by scene-centered components; hence, they proposed a set of global features GIST. Parameters of visual words for text classification tasks such as weighting, dimension, and selection were mapped to image representation in work done by Jun Yang et al. (2007). Visual word vocabulary obtained from feature extractor PCA-SIFT followed by feature selection techniques such as chi-square, mutual information, and point-wise mutual information were utilized to reduce the size of vocabulary and remove unwanted visual words. Bolovinou, Pratikakis, and Perantonis (2013) introduced a bag of spatio-visual word representation obtained from SIFT keypoints. Spherical K-means clustering was used due to the sparsity and large dimensions of spatio-visual word descriptor.

Histogram of visual words was mapped the local features to form a visual vocabulary by Zhou, Zhou, and Hu (2013) and attained an accuracy of about 89.2% for the OT8 dataset. Straightness of edges in indoor as compared to outdoor images, provided the ground for the research proposed in Payne and Singh (2005). Luo and Savakis (2001) used the Bayesian network to combine the semantic information sky or grass detection with low-level features for the classification of indoor and outdoor scenes. Guerin-Dugue and Olivia (2000) use local dominant features and k-nearest neighbor algorithm for classification purpose.

Neural Network Based Techniques

For scene understanding the texture of a particular image is significant. Sorwar, Abraham, and Dooley (2001) purposed DCT based technique to identify texture. DCT feature block followed by classification using EFuNN (Evolving Fuzzy Neural Networks) and ANN (Artficial Neural Netwoks). DCT features along with neural networks happened to successfully classify texture into three classes (bricks, metal, and rural). These low-level and global feature extraction techniques work well when dealing with a small number of classes but when scene classes are increased up to hundreds then these techniques fail in terms of accuracy and time efficiency. Due to the successes of convolutional neural networks (CNNs) in the recent years for processing and classification of image tasks, researchers have moved towards constructing and training their own CNN (Krizhevsky, Sutskever, and Hinton 2012) for scene classification. A large benchmark Places365 (Zhou et al. 2016) was introduced for experimenting with scene classification. The power of convolution neural networks to automatically learn robust features of training data surpassed many handcraft feature extracting techniques. Using this concept Jiangfan Feng, Liu, and Wu (2017) extracted spatial features from CNN and created visual vocabulary-based image representation for the classification of scene images. On MIT indoor dataset achieved an accuracy of 66.09%.

Contrast and spatial information retrieved through the low-level features were combined with high-level feature extraction capability of CNN for salient object detection by Li et al. (2015). Using the capability of low-level features to possess local and global components information along with DCT to enhance the performance of scene detection, we redesign neural networks to incorporate high-level features from the DCT representation of the image and combine them with local and global features. DCT features retain information of image into few coefficients in the frequency domain (Ghosh and Chellappa 2016). Our method proves to be robust for scene recognition as compared to the other deep learning architectures. Table 1 presents a comparison of few papers for scene recognition.

Table 1. Comparison table of scene recognition.

Authors	Method	Accuracy	Classes	Dataset
Raja, Roomi, and Dharmalakshmi (2014)	HSV + DCT with KNN	92.44%	2	IITM- SCID2
Szummer and Picard (1998)	Ohta + MSAR + DCT	90.3%	2	Self Collected
Bolovinou, Pratikakis, and Perantonis (2013)	SIFT + BosVW	91.49%	8	OT8
Zhou et al. (2016)	CNN	55.24%	365	MIT Places
Feng, Liu, and Wu (2017)	CNN + BoVW	66.09%	67	MIT indoor
Our Method	CNN+BOVW+Feature Selection	68.01%	365	MIT Places

Table 2. Effect of SIFT parameter changing.

Parameter Type	*Default parameter*	*Detected number of points*	*Changed parameter*	*Changed number of points*
Contrast threshold	1.0	112	2.0	92
Edge Threshold	20	110	50	83
Sigma	1.0	130	3.0	90

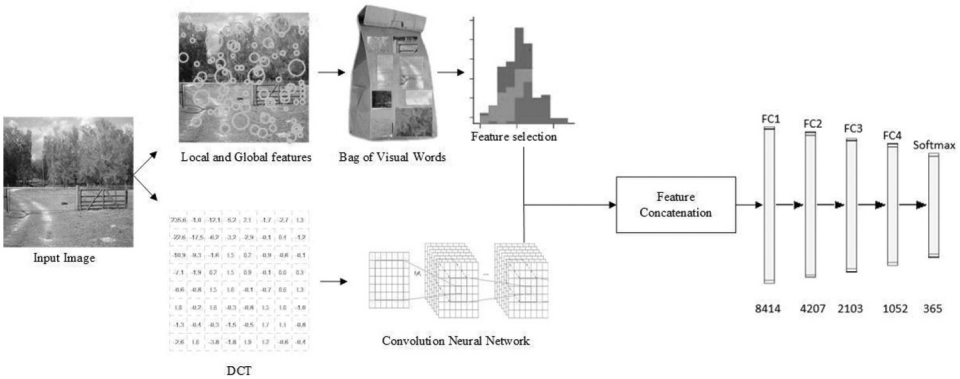

Figure 1. Proposed architecture containing CNN with bag of visual words model.

Proposed Solution

In our proposed solution, we utilized local and global features by extracting Bag of visual word (BOVW) features and texture features from CNN's convolutional layers. We combined them before fully-connected dense layers and classification is done on the final layer using a softmax classifier. The BOVW features are made efficient using feature selection techniques. The proposed solution is depicted in Figure 1 for visual understanding. The details of these techniques are given in the 3.3 Feature selection section whereas the selection and working efficiency are evaluated in Experiments section.

In 3.1 and 3.2 section of Proposed Solution, different types of features extracted from the scene dataset and modeling of the BOVW model from these features are described briefly. In 3.4 section utilization of DCT with CNN model architecture and inclusion of BOVW feature details are explained.

Features for Modeling of Bag of Visual Words

The bag of visual words model (Ghosh and Chellappa 2016) is constructed using two types of features that are SIFT Lowe (2004) and GIST . The main idea is that we represent different types of features from just one feature model that is used in classification. The purpose of the common feature model is to reduce the computational complexity in the testing mode as well as in the training of the Texture CNN model and feature selection methods are easy to apply for pruning less relevant features. We introduce these traditional feature types in DCT-CNN by using this common feature model. Here we have given an overview of above-mentioned features and usage of these features in our method.

SIFT

SIFT finds keypoints by first estimating extremum in scale and space using Difference of Guassian (DOG). The key-points are isolated and contained by removing lower contrast points. The orientation of the key-points are then found using the gradients of image parts. After this step keypoints description is generated using the magnitude and direction of these gradients. This feature extraction technique finds low-level features from the image. We have used this technique for obtaining optimal and efficient keypoints from an image by careful obtaining of parameters shown in Table 3.1.1

GIST

GIST is one of the successful global feature descriptor for image classification with an excellent classification performance rate. It uses a spatial envelope model for the overall abstract representation of a given image. GIST gives properties like openness, naturalness, roughness, ruggedness of an image that is used to classify an image by human as mentioned by Oliva and Torralba (2001). Local features do not provide such kinds of properties of an image to classify it.

We have utilized GIST by dividing the image into four blocks. Each block gives us a 960 feature array and this division of blocks is performed in order to have fixed-size representations of every class. We assumed to have these four feature sets of every image from which the BOVW model is obtained using clustering.

Model Construction

Feature vectors are representations of different parts of an image. We collect representations from three feature descriptor or extractors and concatenate them in a single feature vector.

$$f_i = f_s, f_g \tag{1}$$

where f_s is from SIFT and f_g is from GIST. After the collecting representations of all training images, these representations are clustered through the K-means clustering algorithm Hartigan (1975). The process of the K-means clustering algorithm is initiated at first by defining the total number of cluster centers. The number of clusters defines the BOVW model's output feature vector size. For this, random points are chosen from the descriptions. In the next step, each feature representation f_i is allocated to its neighboring center, by the measure of least Euclidean distance Gower (1985). More formally, if f_c is the collection of centers in a whole set of representations f, then each data point f_i is assigned to a cluster based on

$$\arg\min_{f_c \in f} \; dist(f_c, f_i)^2$$

where dist is that measure of Euclidean distance.

Also, centers f_c are again measured using the mean of all the f_i assigned to their respective clusters.

$$f_c = \frac{1}{|f_i|} \sum_{x \in f_i} x$$

where x is a single feature vector.

This whole process is carried out in the number of iterations until a certain number of iterations. We have defined 1000 iterations for this whole process. This whole process is depicted in Figure 2. Each time when the feature set f_i from the image is given to this clustered model, feature set finds the nearest

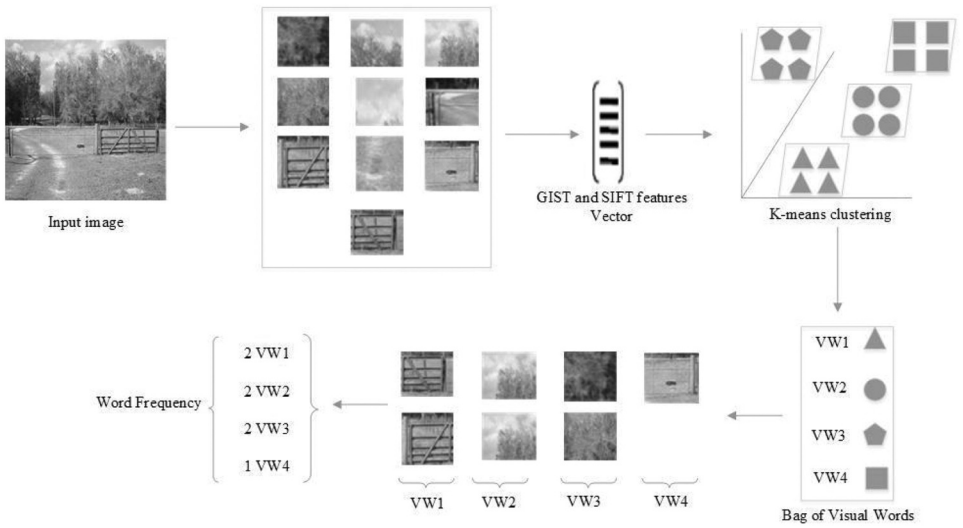

Figure 2. Constructing bag of visual words model.

clusters and the output feature vector is obtained giving the presence of the cluster or visual word in the image. Every number of the vector gives the number of occurrences of the cluster in the description or representation.

Feature Selection

Different feature selection methods are applied to the BOVW model for excluding those features that are less important. For this purpose, we assemble sample data containing all 365 classes and 50 images per class. Feature data from this BOVW model with its labeling class of every image is extracted. Then these feature selection methods are applied in similar fashion like Yang et al. (2007). In the experiment section calculations, experiments, and performance evaluation for pruning the BOVW model are explained in detail. A brief overview of these methods and their usage is explained in this section.

Chi-square Statistics

Chi-square statistics Wilson and Hilferty (1931) has been utilized for the feature selection both in text categorization and image classification methods. It can be used for finding the correlation between two variables. In our case, we are interested to find the correlation between the visual word w and class c of an image. The χ of word w in particular class c is given as

$$\chi^2(w,c) = \frac{(A_o D_o - B_o C_o) * N}{(A_o + C_o)(B_o + D_o)(A_o + B_o)(C_o + D_o)} \tag{2}$$

where N is the total number of training samples.

A_o, B_o, C_o, D_o are the measures which give occurrence of word w and class c both exists, word w exists but class c does not exist, word w does not exist and class c exists and both word w and class c do not exist.

For every chi-square value, the goodness-of-fit test is used to find the level of significance and then this level of significance determines whether this word is suitable for the class or not.

Maximum Entropy

MAXENT (also termed as a conditional exponential classifier) Phillips (2005) classifies image dataset using a maximum entropy modeling framework. This framework deeply reviews experimentally persistent probability distribution and chooses distribution with maximum entropy. If the estimated frequency of occurrence of an image in a class is equal to the actual frequency of occurrence in class then probability distribution is considered experimentally persistent to training samples. Set of weights are used as parameters for the MAXENT classifier to link overlapping regions. The goal is to build a classifier that takes visual words w of an image and generates output value c represented

as (V_w, c) where V_w represents visual information of the image and c is its class. The next step is to demonstrate the training set using empirical probability distribution:

$$p(w, c) = \frac{1}{M} \quad \text{x} \quad \text{total number of times (w, c) occur in corpus} \tag{3}$$

Here M is the size of the training corpus.

$$f(I_i, c) = 0, \text{if} c \neq c \frac{M(I_i, w)}{M(I_i)}, \text{otherwise}$$

Here M(I, w) is the count of number; visual word w occurs in the image I, and M(I) is the count of visual words in I. Weight of a visual word is higher in particular class if its occurrence is relatively higher than other visual words in the same class.

TF-IDF

Tf-IDF Aizawa (2003) measures the importance of the visual word in the Image as well as with the class. Term frequency measures that how frequent the Visual word appears in the class. The tf of the visual word w in particular Image i can be represented as the following:

$$TF_{w,c} = \frac{n_w}{N} \tag{4}$$

where n_w is the number of times a word appears in the class c and N is the total number of words in the Image I. Inverse Document Frequency (IDF) measures that how much important the visual word is for the complete training set. It is possible that a certain visual word is frequent but it may be not important, so this IDF is incorporated with TF to get the quality of the weight. IDF can be represented as

$$DF_{w,I} = \log(\frac{N_d}{N_w}) \tag{5}$$

where N_d is the number of Images in the whole training set and N_w is the number of Images containing this word w. So Tf-Idf term becomes with multiplying both measures to check the relevancy of particular visual word in the class and training set.

$$Tf - IDF = TF_{w,c} * DF_{w,I} \tag{6}$$

DCT-CNN Model

The emergence of deep-learning methods in recent years has improved the classification results. The DCT transform has been used previously with the CNN model by Ulicny and Dahyot (2017). The classification performance is improved and their method is quite effective. We extend this work by using AlexNet with little modifications and BOVW features having local and global features. We have used the AlexNet model for the convolution of DCT images and the BOVW features are included in the fc6 layer. The inclusion at the fc5 layer affects the network's weight adjustment in a way that in forward propagation the prediction and loss are found taking account of both feature vectors. At the time of back-propagation, the weights are adjusted according to the loss of both features used. The network does have some of the properties of the spatial domain as well as in cosine domain. This is the reason for outperforming as well.

Modified AlexNet

For classification of the scene, we used AlexNet network with additional dense fully connected layers. The main reason to use the AlexNet network is that it is considered one of the early deep learning architecture that showed a performance boost in image classification problem. The main focus of this whole method is to show improvement in accuracy by introducing the BOVW model and feature selection in deep architectures. We slightly changed this model by adding two more fully connected layers for the reason of the large size of features other than 4096 standard features by the BOVW feature stream. The architecture diagram of the model is shown in Figure 3. Other architectures may be modified with our technique for performance appraisal. Primarily, the network consists of five convolution layers. Filters are employed over all convolution layers with a stride size of 1.

- Input image of size $224 \times 224 \times 3$ is converted into 1-channel grayscale image before frequency transformation. DCT feature map is computed from grayscale image. A matrix $f(i, j)$ of $M \times N$ dimensions can be transformed into

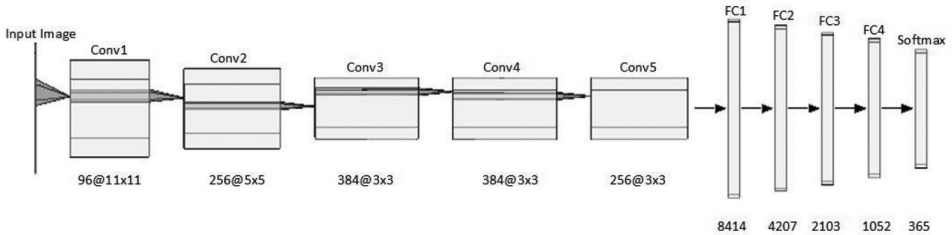

Figure 3. Model diagram of modified AlexNet architecture.

another matrix $f(u,v)$ of $M \times N$ through the definition of Discrete Cosine Transform Ahmed, Natarajan, and Rao (1974) as Equation (7).

$$F(u,v) = \left(\frac{2}{N}\right)^{\frac{1}{2}} \left(\frac{2}{M}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \Lambda(i,j) Cos_{terms} \tag{7}$$

Here $f(i,j)$ represents the matrix before DCT is transformed and $F(u,v)$ denote the matrix after DCT has been performed. Where (i,j) and (u,v) are the coordinates of the both matrices, respectively. Also Cos_{terms} are computed as

$$Cos_{terms} = \cos\left[\frac{\pi u}{2N}(2i+1)\right] \cos\left[\frac{\pi v}{2M}(2j+1)\right] f(i,j) \tag{8}$$

Figure 4 represents the transformed pixels values after applying DCT.

- Input of the first convolution layer of the network is DCT features with size $224 \times 224 \times 1$ having 96 filters of size 11×11 followed by max pooling.
- Output of first convolution layer is fed as an input tothe second convolution layer have 256 filters of size 5×5. The third and the fourth convolution layers have 384 filters of size 3×3. Followed by this last convolution layer comprise of 256 filters of size 5×5.
- In addition to 4096 features extracted from the fifth convolution layer 15,000 features of BoVW are concatenated to fed to the first fully connected layer (FC1). These features are reduced to $9, 532, 4762, 2382, 1192$ features set by passing through FC1, FC2, FC3, and FC4, respectively.
- Softmax function is employed at the end of the network classifies input features to a specific class.

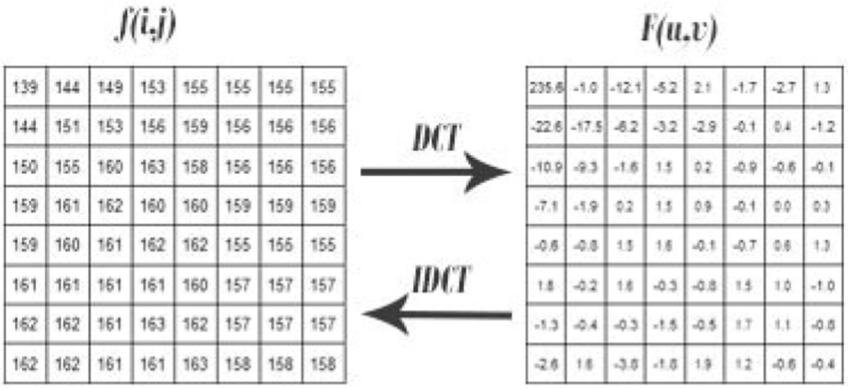

Figure 4. Original data block and data block after DCT transform.

We have employed cross-entropy loss function to measure how good or bad our network is. Followed by a softmax function to obtain probability against each class. Equation (9) refers softmax function:

$$SoftmaxScore : c_m = \frac{e^{c_m}}{\sum_n c_m} \quad \min \{1, 2, \ldots . O\} \tag{9}$$

Here c represents the probability score for classification for o number of classes. When an input image is fed to the network probability score against each class is calcualted, i.e. $c_1, c_2, \ldots . c_o$. Class with the highest score and minimum loss for input sample corresponds to the identified model.

Dataset and Experiments

We have performed a series of experiments to empirically find parameters that are optimum for building the architecture. Appropriate BOVW size and feature selection techniques helped to attain the appropriate additional feature size that is being trained along with DCT-CNN. Also, a comparison of BOVW-based DCT-CNN and past methods are given as proof for improvement of the scene recognition task.

Dataset

We used the standard Places-365 dataset Luo and Savakis (2001) for the classification of scene images. There is a very vast range of scene classes present in this dataset such as urban scenes like train station platform, street, amusement park, etc. Indoor scenes like bedroom, cafeteria, conference center, and the list goes on. The example images are shown in Figure 5. This dataset is chosen for the scene classification task as it contains a very vast range of scene classes. Many deep learning models that are targeting scene recognition problems are using this dataset, and each class contains more than 4000 images. We used all the 365 classes, both training and validation sets are used.

Choosing Suitable BOVW Size

We are using the BOVW model built on SIFT and GIST features. This model gives us a feature set containing local and global representations of the scene. As the BOVW model is given the extracted features and then using k-means clustering model is constructed. The details can be reviewed from section 3.1. In the clustering process, we must specify the total number of clusters to be made. The number of clusters decide the feature size of the input image. In the

Figure 5. Outdoor (a),(b) and indoor (c),(d) scene images from the places365 dataset.

extraction phase, the input image feature set contains an array of features, where each feature gives an occurrence of the visual word.

The size of the BOVW model is important in view of the fact that If the size is too large the feature size increases, it covers more detail but there is a presence of sparsity in the feature set as some classes does not contain some of the visual words and too much mismatch occurs in the feature set. If the size is too small then feature size decreases, sparsity decreases but appropriate detail is not covered resulting decrease in the identification ability of visual words (Wang and Pu 2014) . So we needed to find the appropriate size of the BOVW model which caters above inversely proportional problem. For this purpose we performed an experiment similar to Singhal et al. (180 in which empirically appropriate size is determined using BOVW features with Linear SVM. The accuracy of the BOVW model sizes gives us the reason for using size, except we have used SIFT and GIST features.

In this experiment, we used 50 images of each class from the places-365 dataset for the construction of the BOVW model and training of SVM. Ten images from each class are used to measure the test accuracy of the BOVW-SVM combination. Table 2 shows the effect of SIFT parameter changing for different parameter types, default parameters, detected number of points, changed parameter and changed number of points.

Table 3 shows the training and testing accuracy of the BOVW sizes. Empirically we found the appropriate BOVW size that is 15,000, as the smaller and too much larger BOVW size face less detail or too much sparsity. This method is only used for finding appropriate BOVW size and cannot be used for classification using large datasets. So we have achieved appropriate local and global representation and this representation is used along with the CNN model to cater to the large dataset bottleneck.

Feature Selection Effectiveness

Each word in BOVW has its own significance and its contribution toward representing a particular class. The feature selection methods we are using gives us the relevance of particular visual word with the class. These feature selection methods find relevance based on the level of dependence (chi-square), occurrence in the class and whole set (TF-IDF), and MaxEnt (Prior

Table 3. Effect of BOVW size on accuracy.

BOVW size	Training accuracy	Test accuray
5000	52%	34%
10000	55%	36%
15000	59%	38%
20000	56%	32%
25000	50%	31%

Table 4. Performance of different feature selection method effectiveness on the classification setup shown in the figure. In this experiment, we used the same training set of 365 classes, 50 images each, and 10 images per class are used for testing the classification accuracy used in the above section 4.0.2.

Feature selection methods	Output feature size	Accuracy
BOVW+Chi-square	13403	43%
BOVW+Tf-IDF	13780	41%
BOVW+MaxEnt	13492	42%
BOVW+Chi-square+Tf-IDF	12262	45%
BOVW+Tf-idf+MaxEnt	13180	41%
BOVW+MaxEnt+Chi-square	12733	47%
BOVW+All	11741	39%

probability). Those visual words which are not relevant in more than 150 classes are then pruned to get efficient BOVW representation.

The accuracy shown in Table 4 is the testing the accuracy of the experiment in which we performed classification along with the feature selection method. The table shows that the highest accuracy is achieved by using a combination of chi-square and maximum entropy. Where the combination of all selection methods we have chosen gives a little less accuracy. The reason is that too much pruning of the dataset is done which results in losing the important visual words. We have also seen that Maximum Entropy gives us close results in Tf-IDF and the output size of the BOVW model is also close to each other. Tf-idf and maximum entropy are doing a similar thing that is checking occurrence of the visual word in respective scene classes as well as in the whole training set. But the Maximum Entropy is a probabilistic method that seems to have performed well in this case.

Evaluating Proposed BOVW and DCT-CNN Model

For training our proposed model training is done on the NVIDIA 1080 Ti and the system took approximately 2 days for the complete training of the architecture. Our dataset was divided into training and validation set. The details of the dataset are given in the Dataset section. We performed 80 epochs and in every epoch 8265 iterations carried a batch size of 250 images. We achieved approximately 68% accuracy in the training set and 64% in validation.

The normalized validation loss graph is shown in Figure 6 which shows an abrupt decline in the loss of the architecture. The graph shows continuous decline of the loss with some up and down hops. From 60 epoch the graph showed some stability until the loss becomes stable from 75 to 80 epoch. Training accuracy is 68% and validation accuracy is 64%. The normalized accuracy graph showed in Figure 7 contains two graph lines training and

Figure 6. Training and validation accuracy of the model.

Figure 7. Validation loss of the proposed model.

validation accuracy. Similar to the loss graph the abrupt rise in the accuracy from the start and attaining stability from 60 epoch.

Results and Discussion

The results show considerable improvement in the accuracy by inculcating the improved BOVW model and giving input of DCT to the CNN model. The CNN-BOVW model given by Feng, Liu, and Wu (2017) gave 66% accuracy in

67 classes. Whereas, we achieved 64% accuracy by incorporating feature selection techniques on BOVW and DCT input for efficient texture representation of scenes. Also, the training is carried out using an extensive dataset whereas the dataset used previously was not in really big size.

The AlexNet results on the places365 dataset gave 55.24% accuracy, whereas by incorporating our proposed spatial information method using DCT transform input, BOVW features, and feature selection, the accuracy is greatly improved by 11%.

Our contribution of introducing the features as discussed above in the proposed solution and experiments can also be used to extend more recent Convolutional Neural Networks (CNN) architectures such as VGG16, VGG19, ResNet, and DenseNet. As in the places365 paper (Zhou et al. 2016), these recent architectures are more showing more accuracy than AlexNet architecture. Our focus in this paper is to enhance the classification capability of the CNN models, and we empirically improved using AlexNet as a representative case. The contribution of carefully selecting these local and global spatial features, using DCT-CNN, and fusing it before fully connected layers will further enhance the classification accuracy of these architectures.

Conclusion

This paper presents a scene recognition method through classification by utilizing and synergizing the Bag of visual word (BOVW) features, CNN's convolutional layers, and AlexNet. Bag of visual word (BOVW) features are extracted to utilize local and global features. Texture features are extracted from CNN's convolutional layers. Classification of scenes is carried out using fully connected layers with AlexNet. Local and global features are made efficient by selecting the correct size of Bag of Visual Words (BOVW) and feature selection techniques. The proposed model for scene recognition is also compared with other scene recognition methods, and it outperforms in terms of accuracy on the existing dataset 'places365'.

References

Ahmed, N., T. Natarajan, and K. R. Rao. 1974. Discrete cosine transform. *IEEE transactions on Computers* 100 (1): 90–93.

Aizawa, A. 2003. An information-theoretic perspective of tf–idf measures .*Information Processing & Management* 39(1): 45–65.

Bianco, S., G. Ciocca, C. Cusano, and R. Schettini. 2008. Improving color cnstancy using indoor-outdoor image classification. *IEEE Transactions on Image Processing* 17 (12):2381–92. doi:10.1109/TIP.2008.2006661.

Bolovinou, A., I. Pratikakis, and S. Perantonis. 2013. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition* 46 (3):1039–53. doi:10.1016/j.patcog.2012.07.024.

Feng, J., Y. Liu, and L. Wu. 2017. Bag of visual words model with deep spatial features for geographical scene classification. *Computational Intelligence and Neuroscience* 2017:1–14. doi:10.1155/2017/5169675.

Ghosh, A., and R. Chellappa. 2016. Deep feature extraction in the DCT domain. 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3536–41. Cancun, Mexico. IEEE.

Gower, J. C. 1985. Properties of Euclidean and non-Euclidean distance matrices. L*inear Algebra and its Applications*, 67: 81–97. ISSN 0024–3795. https://doi.org/10.1016/0024-3795(85)90187-9

Guerin-Dugue, A., and A. Olivia. 2000. Classification of scene photographs from local orientations features. *Pattern Recognition Letter* 21:1135.

Hartigan, John A. 1975. *Clustering algorithms*. John Wiley & Sons, Inc.

Kim, W., J. Park, and C. Kim. 2010. A novel method for efficient indoor–outdoor image classification. *Journal of Signal Processing Systems* 61 (3):251–58. doi:10.1007/s11265-009-0446-0.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *In Advances in Neural Information Processing Systems* 1097–105. Lake Tahoe, Nevada, USA.

Lazebnik, S., C. Schmid, and J. Ponce. 2006 June. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), 2169–78. New York, NY, USA. IEEE.

Li, H., H. Lu, Z. Lin, X. Shen, and B. Price. 2015. Lcnn: Low-level feature embedded cnn for salient object detection. *arXiv Preprint arXiv* 1508:03928.

Lowe, David G., 2004. Distinctive image features from scale-invariant keypoints, in the international journal of computer vision, Springer, 60(2): 91–110.

Luo, J., and A. Savakis. 2001. Indoor vs. outdoor classification of consumer photographs using low-level and semantic features. International Conference on Image Processing (ICIP), Thessaloniki, Greece, vol. 2, 745–48.

Ohta, Y., T. Kanade, and T. Sakai. 1980. Color information for region segmentation. *Computer Graphics and Image Processing* 13 (3):222–41. doi:10.1016/0146-664X(80)90047-7.

Oliva, A., and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42 (3):145–75. doi:10.1023/A:1011139631724.

Payne, A., and Singh, S. 2005. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition* 38 (10): 1533–1545.

Phillips, Steven J. 2005. A brief tutorial on Maxent. *AT&T Research* 190 (4): 231–259.

Raja, R., S. M. M. Roomi, and D. Dharmalakshmi. 2014. Classification of scenes into indoor/outdoor. *Research Journal of Applied Sciences, Engineering and Technology* 8 (21):2172–78. doi:10.19026/rjaset.8.1216.

Serrano, N., A. Savakis, and J. Luo. 2002. A computationally efficient approach to indoor/outdoor scene classification. International Conference on Pattern Recognition (ICPR), QC, Canada, vol. 4, 146–49.

Singhal, N., N. Singhal, and V. Kalaichelvi. 2017. Image classification using bag of visual words model with FAST and FREAK. 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT). Coimbatore, IEEE.

Sorwar, G., A. Abraham, and L. S. Dooley. "Texture classification based on DCT and soft computing." The 10th IEEE International Conference on Fuzzy Systems, 2001. Vol. 2. IEEE, Melbourne, VIC, Australia. 2001.

Szummer, M., and R. W. Picard. 1998, January. Indoor-outdoor image classification. In *caivd*, 42. Bombay, India, IEEE.

Ulicny, M., and R. Dahyot. 2017. On using CNN with DCT based image data. Proceedings of the 19th Irish Machine Vision and Image Processing conference IMVIP. Maynooth, Ireland.

Wang, L., and J. Pu. 2014. Image classification algorithm based on sparse coding. *Journal of Multimedia* 9 (1). doi:10.4304/jmm.9.1.114-122.

Wilson, Edwin B., and Margaret M. Hilferty. 1931. The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America 17* (12): 684.

Yang, J., Y. Jiang, A. G. Hauptmann, and C. W. Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. Proceedings of the international workshop on Workshop on multimedia information retrieval. Augsburg, Germany. ACM.

Zhou, B., A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. 2016. Places: An image database for deep scene understanding. *arXiv Preprint arXiv* 1610:02055.

Zhou, L., Z. Zhou, and D. Hu. 2013. Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition* 46 (1):424–33.