Scene Recognition: A Comprehensive Survey

Lin Xie , Feifei Lee , Li Liu , Koji Kotani , Qiu Chen

PII: S0031-3203(20)30011-X
DOI: https://doi.org/10.1016/j.patcog.2020.107205
Reference: PR 107205

To appear in: *Pattern Recognition*

Received date: 17 April 2019
Revised date: 18 December 2019
Accepted date: 11 January 2020

Please cite this article as: Lin Xie , Feifei Lee , Li Liu , Koji Kotani , Qiu Chen , Scene Recognition: A Comprehensive Survey, *Pattern Recognition* (2020), doi: https://doi.org/10.1016/j.patcog.2020.107205

HIGHLIGHTS

- A comprehensive survey on scene recognition is presented.
- Existing scene recognition algorithms are reviewed in the light of feature transformation.
- The relations between various scene recognition algorithms are explored.
- Current benchmarks of different methods are presented and analyzed for comparison.
- Potential problems and future directions are identified.

Scene Recognition: A Comprehensive Survey

Lin Xie[1,a], Feifei Lee[1,a,*], Li Liu[b], Koji Kotani[c], Qiu Chen[d,*]

[a] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China

[b] School of Information Engineering, Nanchang University, China

[c] Department of Electronics and Information Systems, Akita Prefectural University, Japan

[d] Major of Electrical Engineering and Electronics, Graduate School of Engineering, Kogakuin University, Japan

Abstract

With the success of deep learning in the field of computer vision, object recognition has made important breakthroughs, and its recognition accuracy has been drastically improved. However, the performance of scene recognition is still not sufficient to some extent because of complex configurations. Over the past several years, scene recognition algorithms have undergone important evolution as a result of the development of machine learning and Deep Convolutional Neural Networks (DCNN). This paper reviews many of the most popular and effective approaches to scene recognition, which is expected to create benefits for future research and practical applications. We seek to establish relationships among different algorithms and determine the critical components that lead to remarkable performance. Through the analysis of some representative schemes, motivation and insights are identified, which will help to facilitate the design of better recognition architectures. In addition, current available scene datasets and benchmarks are presented for evaluation and comparison. Finally, potential problems and promising directions are highlighted.

Keywords: Scene Recognition, Patch Feature Encoding, Spatial Layout Pattern Learning, Discriminative Region Detection, Convolutional Neural Networks, Deep Learning

1. Introduction

Modern intelligent systems are expected to be capable of coping with different situations without human intervention.

[1] Both authors contributed equally to this work.

[*] Corresponding authors. E-mail addresses: feifeilee@ieee.org, q.chen@ieee.org

It is important for intelligent systems to know the place or context, which helps them understand what might have happened in the past and what may happen in the future. An autonomous mobile robot can be considered as one example of such an intelligence system. It may pass through different places on the way to its destination, so it needs to adjust its actions to accomplish tasks or adapt to the changing surroundings along the way. Adjustment strategies depend on the environment information in the past, present and future. Scene recognition can provide a fundamental description of the content of an image instead of listing the objects in the scene, and it is aimed to help computers understand the environments around them. Scene recognition has been widely used in the applications of human-computer interaction, intelligent robotics, smart video surveillance and autonomous driving. It is also regarded as the prerequisite or the prior knowledge for other advanced computer vision tasks such as image retrieval and object detection. The essential goal for scene recognition is to assign the semantic labels to the given images, these semantic labels are defined by human beings including different natural views, indoor scenes, outdoor environments and etc.

Fig. 1 Image examples for scene recognition. (a) images from *Supermarket* class are shown that scene images may contain diverse objects. (b) images from *Coast* (top row) and *Movie Theater* (bottom row) are used to illustrate the spatial layout variations. (c) images from different categories have semantic ambiguity, they belong to *Church*, *Cloister*, *Library* and *Museum* respectively.

Substantial content, diverse objects, various layouts and semantic ambiguity make scene recognition a more challenging problem compared with other general image classification tasks such as object recognition. Fig. 1 shows some examples to illustrate these challenges. For example, supermarkets often involve plentiful commodities such as various drinks, fruits and foods as shown in Fig. 1 (a). There are several spatial layouts even for some simple scenes such as coast and movie theater shown in Fig. 1 (b). The styles of buildings and decorations in Fig. 1 (c) are identical but they are taken at different scenes, which results in semantic ambiguities between some scenes. Some of these challenges are also present in object detection. However, scene recognition not only concern the existence of objects but also the semantic relations between objects and the contextual information with respect to the background. The large inter-class similarity and intra-class variations are illustrated as in Fig. 2, where the bookstore images highly resemble the library images and the presented fountain images are totally different. In addition, it is also difficulty to precisely represent the complicated semantics and spatial structural information for scene recognition.

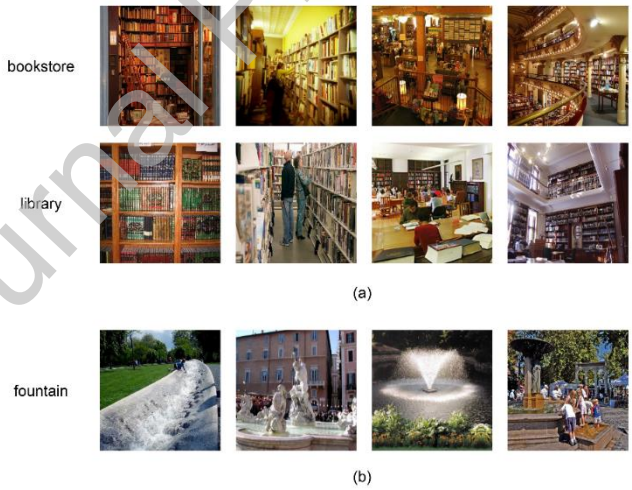

Fig. 2 Illustration of the inter-class similarity and intra-class variations. (a) images from *Bookstore* (top row) extremely resemble the images from *Library* (bottom row). (b) images from *Fountain* are very different.

According to the manipulation of features extracted from images, scene recognition algorithms can be roughly grouped into the following six major categories: global attribute descriptors, patch feature encoding, spatial layout pattern learning, discriminative region detection, object correlation analysis and hybrid deep models, as shown in Fig. 3.

In the early 2000s, scene image representations mainly relied on ***Global Attribute Descriptors***, which is constructed by some low-level visual properties to model the perception of human beings. Typical global attribute descriptors include GIST [1], Semantic Typicality [2], Edge Straightness Analysis [3], CENsus TRansform hISTogram (CENTRIST) [4], Local Difference Binary Pattern (LDBP) [5] and multi-channel CENTRIST (mCENTRIST) [6]. The contribution of global attribute descriptors to scene recognition is limited due to the complex visual constitutions of scene images.

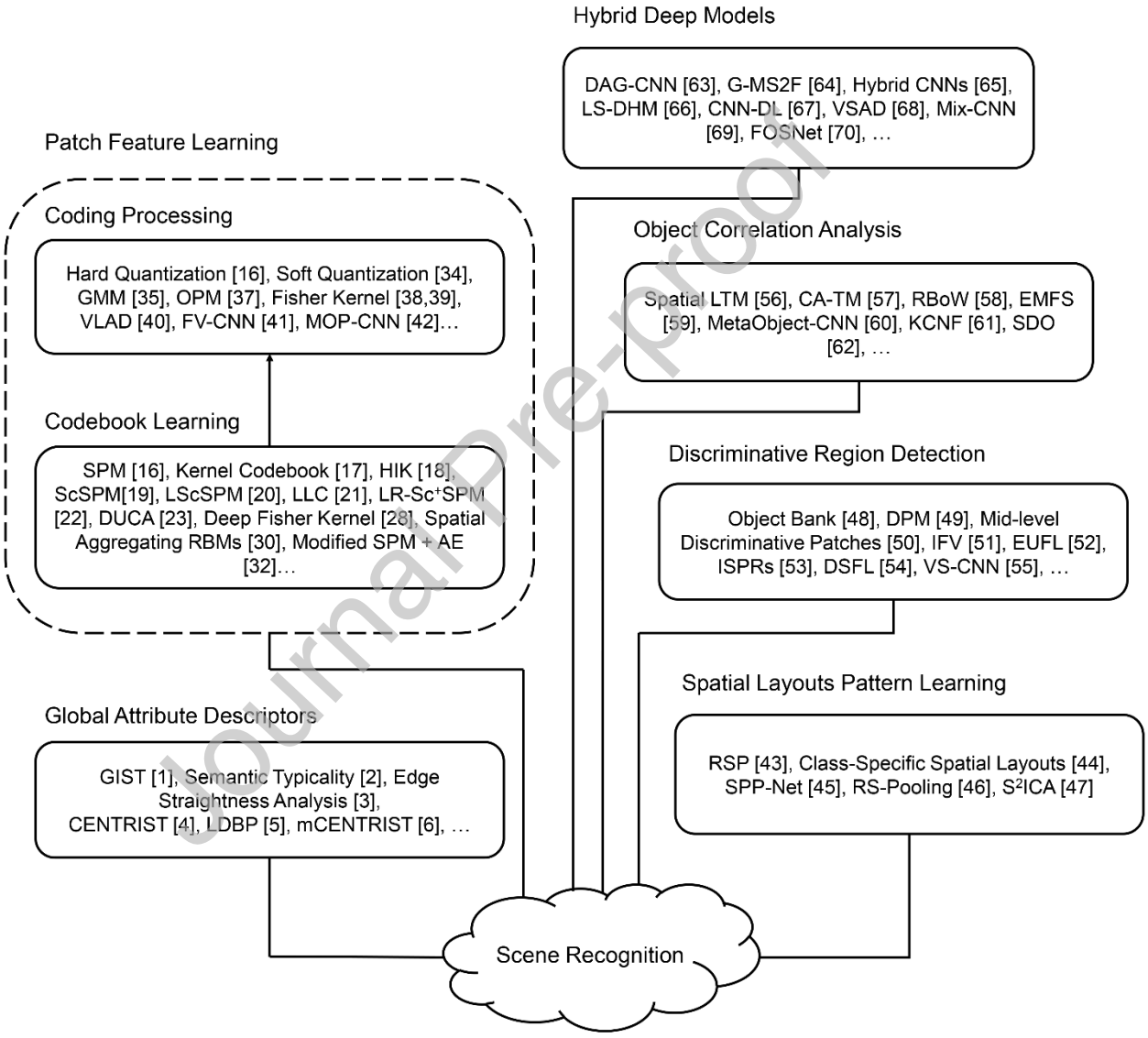

Fig. 3 The taxonomy of scene recognition algorithms. All of them are roughly grouped into the six major categories in this paper.

To improve the recognition performance, many researchers shifted their focus to ***Patch Feature Encoding***. Some notable local visual descriptors have been widely used in the patch feature extraction including Local Binary Patterns (LBP)

[7], Scale Invariant Feature Transform (SIFT) [8], Histogram of Oriented Gradients (HOG) [9], Speeded Up Robust Features (SURF) [10] and Oriented Texture Curves (OTC) [11]. Bag-of-Visual-Words (BoVW) [12,13,14] framework is introduced to integrate a large number of local visual descriptors into an image representation. Spatial Pyramid Matching (SPM) [15, 16] is a standard component proposed for the BoVW framework to compensate for the missing spatial structural information. Specifically, local visual descriptors are quantified into finite visual words, and then the entire scene image is described as the occurrence frequency of visual words. The former is defined as the codebook learning and the latter is defined as the coding processing in this paper.

The quality of the learned codebook has a great impact on the recognition performance. The kernel codebook [17] is proposed to exploit the ambiguity between visual words and reaped higher benefits in a high-dimensional feature space. The histogram kernel [18] is introduced into K-means to lead to more effective codebook and higher recognition accuracy. It has been found that conventional clustering algorithms in codebook learning are sensitive to outliers and result in large quantization errors. To overcome this problem, many codebook learning algorithms based on the reconstruction of inputs have been adopted in scene recognition. Sparse coding [19] and its derivate algorithms [20, 21, 22] are developed to adaptively learn the implicit codebook and reduce the quantization errors. Recently, codebook is learned from scene representative patches using sparse linear coding in Deep Un-structured Convolutional Activations (DUCA) [23]. Moreover, extra information is incorporated into visual words to form a more powerful codebook. For example, contextual information and the relationships between visual words are exploited to learn the codebook in [24]; the divisive information theoretic feature clustering is used to learn a compact codebook by combining multiple kinds of features in [25]; local features are extracted from multi-resolution images to learn the codebook with more spatial information in [26]; typical spatial configurations and shapes occurring across the objects subparts at different scales are used to build visual words in [27]. Moreover, some novel deep architectures stem from conventional methods are proposed to encode local patch features such as deep fisher kernels [28] and manifold regularized deep architectures [29]. Deep neural networks can also be

considered as a special feature encoding method, the codebook corresponds to the learned network parameters. Restricted Boltzmann Machines (RBM) [30] and autoencoders [31, 32] have been exploited in the BoVW framework.

Given the learned codebook, coding processing helps transform local visual words into the image representation. One of the widely used methods of coding processing is hard quantization, in which each local visual word is assigned to one visual word. Soft quantization [33, 34] and Gaussian Mixture Model (GMM) [35] are used to cope with the local visual descriptors that resemble multiple visual words. To integrate spatial information into image representations, SPM is devised to indicate the variations of the regional visual words distributions. Similarly, multiscale spatial pooling is proposed to combine with sparse coding in Sparse coding Spatial Pyramid Matching (ScSPM) [19]. It has been regarded as the standard module in combination with the codebook learning algorithms based on the reconstruction of inputs. In addition, there have been some works concerning the construction of spatial pyramids. The discriminative spatial pyramid [36] is proposed to automatically select the weights of all pyramid levels to maximize a discriminative power. Orientational Pyramid Matching (OPM) [37] utilized the 3D orientations to index the patches in the orientational space instead of the positions of local patches to form the pyramid. Recently, Fisher kernel [38, 39] and Vector of Locally Aggregated Descriptors (VLAD) [40] have been utilized to aggregated the patch convolutional features in Fisher Vector pooling (FV-CNN) [41] and Multi-scale orderless pooling (MOP-CNN) [42] respectively. Both of them have achieved impressive performance.

Spatial Layout Patterns Learning is another useful method for scene recognition. There are always some underlying spatial layouts for some specific scenes, which can be utilized to promote the recognition performance. Randomized Spatial Partition (RSP) [43] treats the optimal spatial partition for each category and classification task as a joint optimization problem. Subsequently, class-specific spatial layouts [44] are discovered from the spatial partition on convolutional feature maps. The design principles of some customized end-to-end networks also concentrate on capturing some useful spatial layout patterns. Typical customized modules for spatial structures include Spatial Pyramid Pooling (SPP) [45], Randomized Spatial pooling [46] (RS-pooling) and spatial unstructured layer [47].

Discriminative Region Detection is aimed at autonomously selecting some crucial regions for scene recognition. Object Bank [48] and deformable part-based models [49] resort to object detection algorithms to gain the discriminative regions. Whereas other methods attempt to identify the discriminative regions from a large number of image patches, such as unsupervised discriminative clustering [50], entropy-rank curves [51] and density estimation [52]. To suppress the noisy features, learning important spatial pooling regions (ISPRs) [53] utilizes part filters to retain the response of important regions. To overcome the inner-class variance and interclass similarity, Discriminative and Shareable Feature Learning (DSFL) [54] is developed to enforce the learned features from the same category to be close to each other and the learned features from different categories to be far away from each other. By feeding the pre-trained CNNs with the scene images enhanced by detected saliency, the obtained deep visually sensitive features [55] are proved to be effective for scene recognition.

Object Correlation Analysis seeks to model the relations between the distribution of diverse objects and scene categories. Early explorations of object correlation are built on topic models in which object recognition is a prerequisite. Typical topic models include spatially coherent latent topic model [56], context aware topic model [57], reconfigurable models [58]. In light of the objects distribution between different scenes, some models are proposed to exploit the co-occurrence patterns of each category such as the context model in the semantic manifold [59], MetaObject-CNN [60], Kernel Co-occurrence Noise Filter (KCNF) [61] and Semantic Descriptor with Objectness (SDO) [62].

Recently, ***Hybrid Deep Models*** have been proved to be one of the most effective methods for scene recognition. On the one hand, the intermediate layers in CNNs often capture local features while the top layers extract holistic features because of the extensive receptive fields. Multi-stages convolutional features can be taken into account in the end-to-end networks such as directed acyclic graph CNNs (DAG-CNNs) [63]. It is more convenient to directly fuse or transform multi-stages convolutional features after they are extracted from pre-trained models [64, 65, 66]. CNN-DL [67] attempts to embed sparse coding in the customized end-to-end networks to transform convolutional features for scene recognition. On the

other hand, the Scene-CNN and Object-CNN can be considered to extract holistic and local features for scene recognition. For example, the holistic features extracted from the Scene-PatchNet and the local features extracted from the Object-PatchNet are combined for scene recognition by the Vector of Semantically Aggregated Descriptors (VSAD) [68]. The mixed datasets of objects and scenes are constructed to train the Mix-CNN [69] to learn the codebook, which can be shared by Object-CNN, Scene-CNN and Mix-CNN in feature encoding to obtain the ultimate image representation. FOSNet [70] is another end-to-end CNN framework based on the fusion of object and scene information, which assumes that the adjacent patches of a single image belong to the same scene class.

Despite extensive literatures in scene recognition, there are few literatures that present a comprehensive review on this fundamental field. Some related reviews and surveys have been published, their main contribution and limitations are presented as follows:

1) X. Wei et al. [71] provides a comprehensive review on the visual descriptors for scene recognition. An empirical study to assess these visual descriptors is also conducted. This work especially focuses on the unsupervised feature extraction methods and considers the feature extraction and the classifier training as two independent stages. Few deep learning models are mentioned and various novel deep models for scene recognition are also not covered by this work.

2) E. Anu et al. [72] enumerates various scene recognition methods that attempt to build the intermediate semantic representation to reduce the semantic gap between human beings and computers on scene understanding. The principles and limitations of these methods are investigated and discussed. However, the mentioned methods are not grouped into several explicit categories, and the difference and relations between them have not been explored. Deep learning based methods are also not included in this work.

3) V. Singh et al. [73] presents a brief review on various early methods for scene recognition. These methods are grouped into three major categories: methods based on object detection, methods based on low-level image features and other methods. Many methods listed in this review are constrained to identify extremely particular scenes with few

categories such as city against suburb, indoor against outdoor and several simple natural scenes. Some prevailing feature encoding methods and deep learning based models are not covered.

Different from these reviews mentioned above, the major contributions of this paper can be summarized as follows: First, we provide a more comprehensive and systematic overview of scene recognition algorithms, including global attribute descriptors, patch feature encoding, spatial layout pattern learning, discriminative region detection, object correlation analysis and hybrid deep models. Second, we compare and analyze different levels of scene recognition algorithms. By examining the relations between the algorithms, some superiorities and potential problems are uncovered. Third, the prevailing databases for scene recognition are presented for evaluating the recognition performance of various algorithms. Finally, based on these findings, the promising directions of future research are identified for beginners. From the perspective of practitioners in this field, our work can provide some useful insights and facilitate the design of better scene recognition architectures.

The remainder of this survey is structured as follows. It starts with the general pipeline for image classification. After that, various kinds of scene recognition algorithms are demonstrated as the following order: global attribute descriptors, patch feature encoding, spatial layout pattern learning, discriminative region detection, object correlation analysis and hybrid deep models. Some representative works of each kind of algorithms are illustrated and explained. In addition, we introduce the popular databases for scene recognition and compare the recognition performance of different levels of algorithms. Finally, we summarize the contents of this paper and concisely discuss the future prospects of scene recognition.

2. General Pipeline for Image Classification

As shown in Fig. 4, the general pipeline of most alternatives for scene recognition consists of three vital stages. Almost without exception, given an input image, the first step is feature extraction. After acquiring the visual characteristics, feature transformation is usually adopted to capture some scene traits to form the image representation. Finally, classification is conducted using the learned classifier. In the early image classification algorithms, visual characteristics are represented

by handcrafted features such as edges, corners and various local visual descriptors. With the advent of large scale image datasets and the breakthrough of deep convolutional neural networks (DCNN), these handcrafted features are gradually replaced by the deep features because their stronger expressive power. Feature transformation is the critical technique in scene recognition algorithms, which is the core topic of this review. Some conventional methods can be used for the final classification such as SVM and softmax regression. Given the image representation, the recognition performance between diverse classifiers has little difference. It is worth noting that some end-to-end deep learning networks unify feature extraction, feature transformation and image classification into a coherent pipeline.

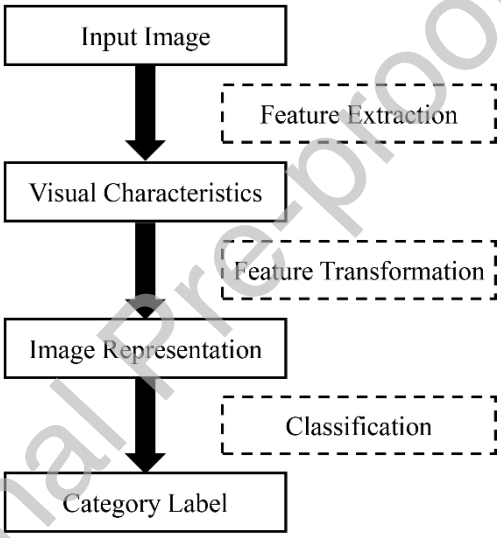

Fig. 4 General pipeline for image classification.

3. Review of Scene Recognition Algorithms

3.1 Global Attribute Descriptors

Global attribute descriptors are devised to identify some particular visual attributes such as texture and geometric pattern. Scene recognition requires global visual descriptors to possess properties that are different from those in object recognition domains. GIST [1] is a special visual representation of scenes in which a set of perceptual properties are defined to identify the semantic meanings for different scenes. It discards the routine that recognizes scenes based on the segmentation and processing of individual objects or regions. The scene representation is modeled by five spatial envelope

properties, including the degree of naturalness, openness, roughness, expansion and ruggedness. Although this low-level descriptor builds a direct interface between visual perception and semantic knowledge, the deficiency of limited information results in poor performance for complex scenes. The Semantic Typicality [2] is a function of the occurrence frequency of scene category attributes, and it is used to measure the similarity of natural real-world scenes with respect to six scenes including coast, rivers/lakes, forests, plains, mountains and sky/clouds. CENTRIST [3] emerged as an important holistic image feature with stronger generalizability for scene recognition. It mainly encodes the structural properties and suppresses detailed textual information. Compared with GIST, the recognition accuracy of CENTRIST for indoor categories is much higher, and some rough geometrical information may account for this result. However, there are also several factors that may reduce the recognition performance; for example, it is not invariant to rotations and ignores useful information in the color space.

3.2 Patch Feature Encoding

3.2.1 Pipeline of Bag of Visual Words Framework

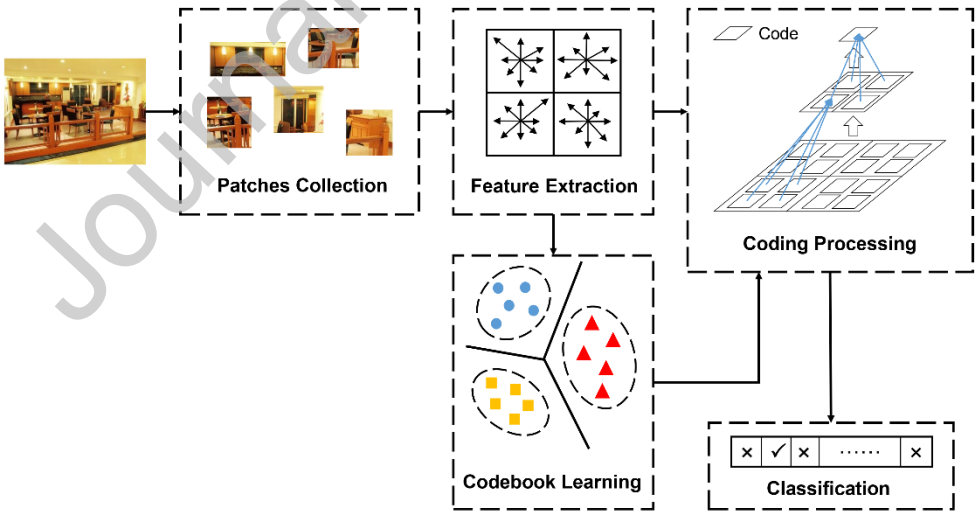

Fig. 5 General pipeline of Bag of Visual Words (BoVW). Firstly, local visual descriptors are extracted from the densely sampled image patches. Secondly, these local visual descriptors are used to learn the codebook. Thirdly, local visual descriptors are aggregated into the image representation by the coding processing. Finally, the classifier is training for scene classification.

Before the rise of deep learning, BoVW framework is a kind of the most important feature transformation and extensively used for image classification. It is introduced from text analysis [12] and employed to integrate a large number of local visual descriptors into the image representation [13]. As shown in Fig. 5, BoVW framework consists of three basic modules: *local feature extraction*, *codebook learning* and *coding processing*. Given an image as the input, the local visual descriptors are usually extracted from many densely sampled image patches. The sliding sampling window has a fixed size, so the extracted local visual descriptors have the same dimensions, which is convenient for the subsequent codebook learning. Two spatially adjacent patches are overlapped so that the successive variations or uniform regions can be captured as much as possible. It has been shown that the extracted local features can provide more powerful clues and are more robust to occlusion and spatial variations than global visual descriptors [14]. Codebook learning is aimed at distilling some visual words that can effectively represent existing local visual descriptors. The primary method of codebook learning is K-means clustering, which partitions local visual descriptors into several clusters according to the Euclidean distance between them. Each cluster is a set of local visual descriptors with similar characteristics. The center point of each cluster is regarded as the unique visual word, and all visual words form the codebook. A local visual descriptor is assigned to one of the visual words by its nearest neighbor; consequently, an image can be represented by the distribution of these visual words. After codebook learning, coding processing utilizes the learned visual words to depict the contents of the entire image. For simplicity, the number of occurrences of all visual words can be collected as the ultimate image representation, which is named as hard quantization. We will go through other coding processing variants. Consequently, many local visual descriptors scattered in the whole image are aggregated into a single distribution vector. BoVW models the distribution of different local visual descriptors, which may be extremely significant for the separation of most scenes because different scenes may consist of discrepant objects and backgrounds.

Robust local visual descriptors are important for BoVW. Outstanding local visual descriptors include LBP [7], SIFT [8], HOG [9] and SURF [10]. However, their original purpose is to depict edges or corners without considering the

characteristics of scene images. OTC [11] is developed to capture the particular local characteristics. Multiple curves are constructed to represent the color variation along different orientations. The texture of the image patch is characterized by the shapes of these curves, which has proven robust for illuminating differences and geometric distortions. To process homogeneous patches, an H-bin normalization scheme is proposed to avoid the creation of false features. OTC has considerably improved recognition accuracy over other local visual descriptors and has shown potential in conjunction with extra complementary features.

3.2.2 Codebook Learning

Codebook learning is to obtain some representative visual words that can reflect some intrinsic characteristics from a set of visual features. As is mentioned above, the learned codebook using K-means clustering is composed of the cluster centers, which are considered to be crude because the outliers can easily result in a completely different codebook. One significant factor in codebook learning is the similarity measurement between local descriptors and visual words. It has been found that the histogram intersection kernel is more effective than the Euclidean distance in regard to histogram features [17]. To speed up codebook learning and exploit the advantage of the histogram intersection kernel simultaneously, the kernel K-means algorithm is proposed [18]. The similarity between the local descriptors and visual words is implicitly calculated in the kernel space. However, the costs of marginal gains are the complex training procedures and increased computation requirements. In addition, to make the learned visual words more robust and representative, reconstructing the original visual features by the codebook has been widely adopted in scene recognition.

For consistency, the primary codebook learning is equivalent to the following optimization process:

$$\min_{\mathrm{U,V}} \sum_{n=1}^{N} \|x_n - u_n \mathrm{V}\|^2 \quad \text{subject to } Card(u_n) = 1, |u_n| = 1, u_n \geq 0, \forall n \tag{1}$$

where $\mathrm{V} = [v_1, \ldots, v_K]^T$ is the codebook with K visual words to be found, u_n is the weight indicator of the local descriptor x_n on all visual words, $Card(u_n) = 1$ means that only one element of u_n is nonzero, and $|u_n|$ is the $L1$-

norm of u_n. It is obvious that the constraints of this optimization are too restrictive to reach the appropriate minimum. By relaxing the constraints, a more precise codebook can be learned from the inputs using sparse coding:

$$\min_{\mathrm{U,\,V}} \sum_{n=1}^{N} \|x_n - u_n \mathrm{V}\|^2 + \lambda |u_n| \quad \text{subject to } \|v_k\| \le 1, \forall k = 1,2,\dots,K \tag{2}$$

where only an L2-norm constraint on $\|v_k\|$ is applied to avoid trivial solutions, and $|u_n|$ induces the sparsity. In the training phase, all local features are used to solve the overcomplete codebook; in the coding phase, the codes are inferenced by reconstructing the individual local descriptor using several visual words.

Sparse coding is firstly introduced in ScSPM [19]. After that, many different codebook learning schemes have been developed to reduce the quantization errors and make the inferred codes more separable. Laplacian Sparse coding (LScSPM) [20] is proposed to exploit the dependence among local features, which alleviated the problem that the attributes of local features and inferred codes are inconsistent. In LScSPM, the optimization problem of sparse coding is reformulated by incorporating the Laplacian matrix. This regularization on sparse coding helps to significantly reduce the quantization error and maximally preserve the similarity between local features, which leads to better performance on scene recognition, but the Laplacian matrix also increases the computational burden. Considering the tradeoff between the performance gains and lower computational overhead, another codebook learning strategy called LLC (Locality-constrained Linear Coding) [21] is proposed to improve the quantization process. The locality regularization is emphasized in LLC instead of sparsity. The main idea of LLC is to select several visual words that resemble the local feature, the locality regularization is proportional to their similarity, and the code is derived from the local coordinate system based on the selected visual words. One of the merits of LLC is its high efficiency even with a large number of visual words because it has an analytical solution and a fast approximation. Motivated by the observation that there are common items and specific items in images from the same category, the low-rank and sparse matrix recovery technique [22] is proposed to incorporate common and specific attributes into the codebook. This technique combines the codebook learning strategies in ScSPM and LLC, the final codes show

stronger robustness for the variations in the same class. In these efforts described above, the correlation in the original feature space is preserved, and the variation in the same class is also considered. Recently, the local visual descriptors are replaced by convolutional features to represent image patches, which leads to the improvement of recognition accuracy. For example, some scene representative patches are collected to learn the codebook using sparse linear coding in DUCA [23].

3.2.3 Coding Processing

The conventional K-means clustering used in BoVW is hard quantization, which means that each local visual descriptor is assigned to only one visual word. The histogram of the frequencies for all visual words h is calculated according to the closest visual word v to the local visual descriptor x as follows:

$$h(i) = \begin{cases} 1, if\ i = arg\min_j(\|x - v_j\|) \\ 0, otherwise \end{cases}, i = 1,2,\dots,K \tag{3}$$

Regarding the points located on the boundary of two or various clusters, Euclidean distance leads to confusing results. Soft quantization [33, 34] allows us to bypass this difficulty and assigns each local visual descriptor to more than one visual word. The probability density over visual words is estimated by the Gaussian kernel, so the histogram h is

$$h(i) = \frac{\exp(\|x - v_i\|^2/\sigma)}{\sum_{j=1}^{K}\exp(\|x - v_j\|^2/\sigma)}, i = 1,2,\dots,K \tag{4}$$

where σ is a smooth parameter. This approach not only makes the relevant visual words ambiguous but also defines the uncertainty for the histogram of visual words. It has been shown that this approach can decrease the quantization errors in coding processing and increase the recognition performance. The main drawback is that the assumption of a Gaussian distribution for a histogram of visual words rarely holds.

In contrast to the general object recognition, BoVW is insufficient for scene recognition because it merely detects whether a visual word exists or the frequency of a visual word and disregards the spatial arrangement of local visual features. To overcome this problem, SPM [16] is proposed to incorporate spatial information into the ultimate image representation and achieved significant success. SPM generalizes the BoVW paradigm into spatial regions with different scales, and defines the matching rules for the final recognition. The spatial pyramid is built on the original image from coarse-grained spatial partitions to fine-grained spatial partitions. The contributions to recognition of the spatial regions with different sizes are considered to be distinct according to the spatial pyramid matching kernel. There have been many works that attempt to further improve some steps in SPM. For the weights of different scales, the work in [36] gives another method for automatically selecting weights to maximize the discriminative power.

In addition to hard quantization and soft quantization used in primary codebook learning schemes, multiscale spatial pooling is another important coding processing technique in combination with other codebook learning schemes, which is firstly applied in ScSPM. The derived codes are aggregated into the ultimate image representation by means of a series of progressive pooling operations. The dimensionality of the ultimate image representation is equal to the number of visual words in the learned codebook. Instead of concatenating all distribution vectors from different spatial regions, multiscale spatial pooling aggregates many derived codes across different scales into the ultimate image representation with lower dimensionality, which substantially reduces memory consumption. There have been many kinds of pooling operations, such as maxpooling and average pooling. A comprehensive cross-evaluation of several types of codebook learning and pooling schemes is presented in [34].

Fisher kernel [38, 39] and Vector of Locally Aggregated Descriptors (VLAD) [40] are two methods that have been widely used to aggregate local visual descriptors into the image representation in general image classification. They can also be applied to the patch features out of CNNs, and the effectiveness of them has been verified in Fisher Vector pooling (FV-CNN) [41] and Multi-scale orderless pooling (MOP-CNN) [42] respectively.

3.2.4 Feature Encoding Using Deep Networks

Before the prosperity of CNNs, there have been some deep neural networks such as deep belief networks, deep Boltzmann machines and deep autoencoders. However, the sole network for scene recognition often suffers from the notorious training process and results in poor performance. Therefore, they are utilized to transform some low-level visual features into more abstract image representations by the reconstruction of inputs. It has been shown that the trivial reconstruction cannot lead to more powerful image representations, so the feature transformation is constrained in some feature space with specific properties. The learned parameters of these deep networks can be considered as the implicit codebook, the coding processing methods mentioned above are still available for these algorithms. Some novel deep architectures stem from conventional methods. For example, SVMs with Fisher kernel [28] for scene recognition can be interpreted as deep networks to encode patch features; a manifold regularized deep architecture [29] is proposed to capture the nonlinear structure of the local visual descriptors by incorporating the sparse regularizer into the kernel manifold space. Some fundamental neural networks are also exploited for feature encoding. Spatial aggregating RBMs are first utilized to encode local visual descriptors in [30], where the RBM is regularized to fit both sparse and selective distributions, which led to more competitive results. Similar regularization is added to the learning process of the AE network in [31]. The work in [32] attempts to encode local visual descriptors using various AEs, and it integrated some spatial structure information into the image representation.

3.3. Spatial Layout Pattern Learning

Spatial layouts can be different for various classes of scenes, so class-general spatial layouts may not fit different classes well and reduce the discriminative power. It has been proved that learning class-specific patterns of spatial layouts is an effective alternative. To acquire the optimal spatial structure information for scene recognition, RSP [43] based on SPM attempted to mine the most descriptive image layout pattern for each class. From the viewpoint of local regions, RSP can well separate the distinguishing components in the scene. Specifically, the best pattern with minimum validation error

is selected as the optimal description of spatial layout information for each class, and the ultimate multiclass classifier is recast by training multiple binary classifiers using the selected best patterns for each class. As presented in Fig. 6, the single spatial pattern may not be sufficient to describe the elastic spatial layout; a sequence of patterns weighted in proportion to their discriminative power is combined, and the selected patterns and related weights are progressively determined by the boosting algorithm. This method adopts multiple spatial patterns to describe the specific complicated spatial layout, which makes it possible to process some outliers.

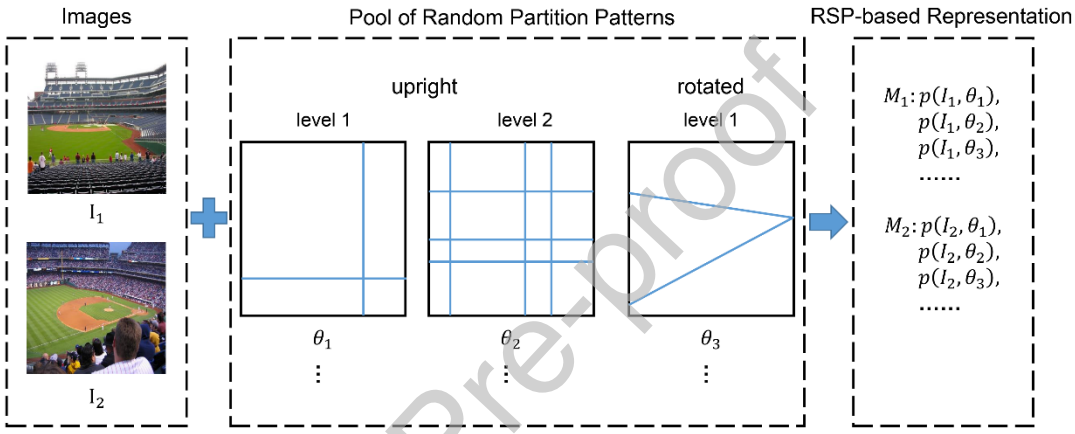

Fig. 6 The image space is randomly partitioned into sub-regions of various sizes and shapes by the learnable parameters, each partition pattern can be represented as a histogram feature $p(l_i, \theta_j)$, where i denotes the level, j denotes the partition patterns. (reproduced from [43]).

Similarly, the work in [44] extended RSP on the feature maps of CNNs to discover class-specific spatial layouts for scene recognition. To seek various class-specific layouts as much as possible, a spatial pyramid with L levels is built on the feature maps. This spatial pyramid is different from the counterpart in SPM. As shown in Fig. 7, the feature maps are randomly divided into a fixed number of subregions of various sizes subject to the uniform distribution, where the fixed number is proportional to the level, each spatial pattern corresponds to the stacked feature vectors of all regions. To obtain adequate random spatial patterns, the random partition process is repeated multiple times, which derives many different spatial partition results for an image. To overcome the large variations in a particular scene class, a linear combination of all spatial partition results is performed to construct the best spatial layout. Finally, the l_1-regularized max-margin

formulation is proposed to discover the class-specific spatial layouts. The weight coefficients of all spatial partition schemes and the binary classifiers for each category are simultaneously learned from the formulation. Experiments have shown that this joint optimization can lead to competitive results.

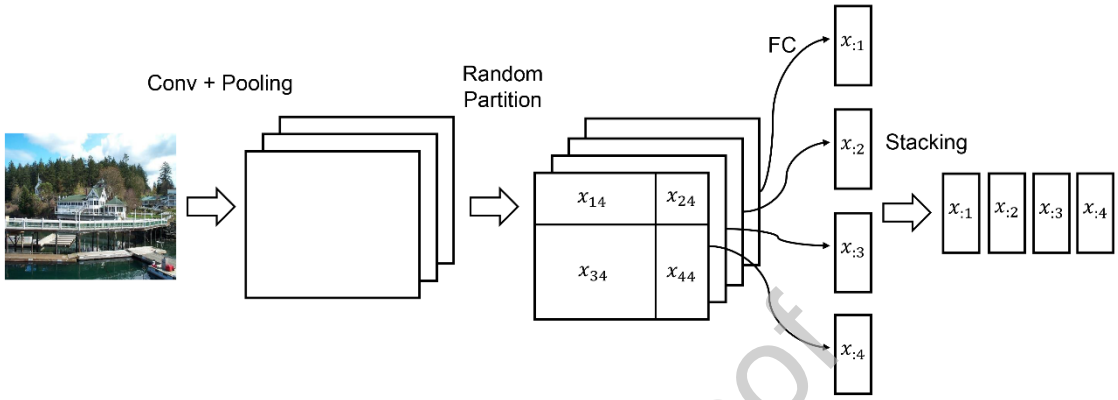

Fig. 7 Spatial partition scheme for the discovery of class-specific spatial layouts. Multiple random partitions on the 2D feature maps of CNNs form different spatial layouts. The features from different partitions are stacked as the image representation. (reproduced from [44]).

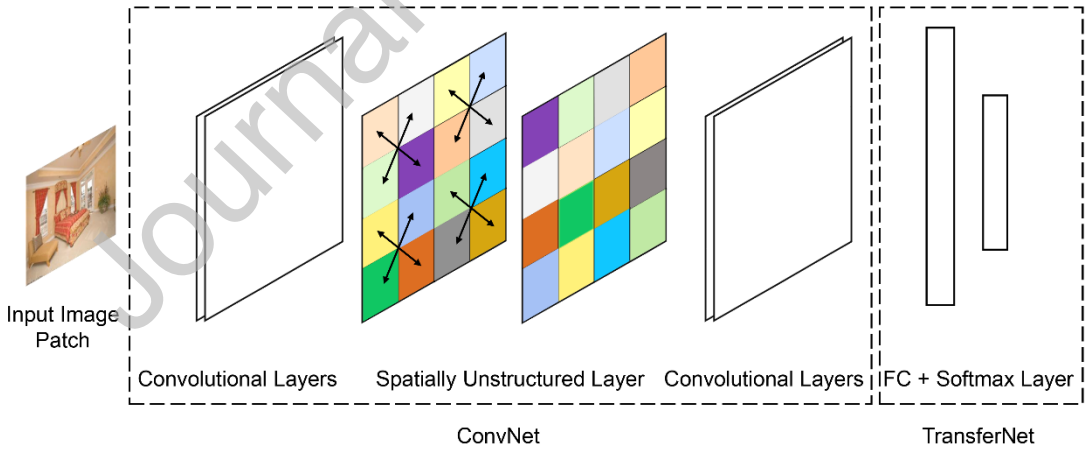

Fig. 8 The spatial unstructured layer is incorporated into CNNs (reproduced from [47]). The feature maps are split into a specified number of blocks. The local swapping operations are performed on each block, which makes the convolutional features robust to local spatial deformation.

Many notable CNNs structures developed for the object recognition task [74] (e.g. ImageNet challenge) may not be

sufficient for scene recognition. Some customized network structures have been proposed to capture more spatial structure information. SPP-Net [45] replaces the last pooling layer of CNNs with the spatial pyramid pooling layer. The feature map is partitioned into several spatial bins. These spatial bins have sizes proportional to the image size, so the number of bins for the images with arbitrary sizes are the same. For each bin, max pooling is conducted on the responses of each filter. As a result, the output of SPP-Net is the fixed-dimensional vector that maintains some useful spatial structure information. To deal with the large-scale deformations and scale variations in the spatial layouts of scene images, the spatial unstructured layer [47] is incorporated into CNNs. As presented in Fig. 8, the feature maps of the first subsampling layer are split into a specified number of blocks. The local swapping operations are performed on each block, which breaks the original spatial order to some extent. It is intended to obtain robust feature representations to address spatial deformations. By means of transfer learning, the two-deep convolutional architectures with and without the spatial unstructured layer are adapted to scene recognition. Both convolutional activations are concatenated to form the final image representation. Although the image representation may be robust to spatial deformation, the discriminative details can be diluted by the local swapping operations.

In contrast to the predefined spatial partition schemes on feature maps, many researchers have attempted to devise adaptive pooling layers to examine more flexible spatial structures. Analogous to the practice in RSP, the RS-pooling layer [46] is proposed to incorporate the appropriate spatial layout information into CNNs. Specifically, according to a set of randomized spatial partition patterns, the feature maps are divided into a fixed number of subregions with various sizes and shapes, and these subregions are further partitioned into a fixed number of cells. The max-pooling on each cell results in a fixed-dimensional vector as the input of the fully connected (FC) layer. The optimal image layout description for the input image can be determined by a maxout objective function. The architecture of RS-pooling CNNs is shown in Fig. 9. The optimal image layout pattern is chosen by the maximum response of FC layers for all random spatial patterns. The whole training process involves only randomized spatial pooling layers, subsequent FC layers and the final maxout objective

function. Therefore, the number of the training parameters and the complexity of this model depend on the number of random spatial patterns.

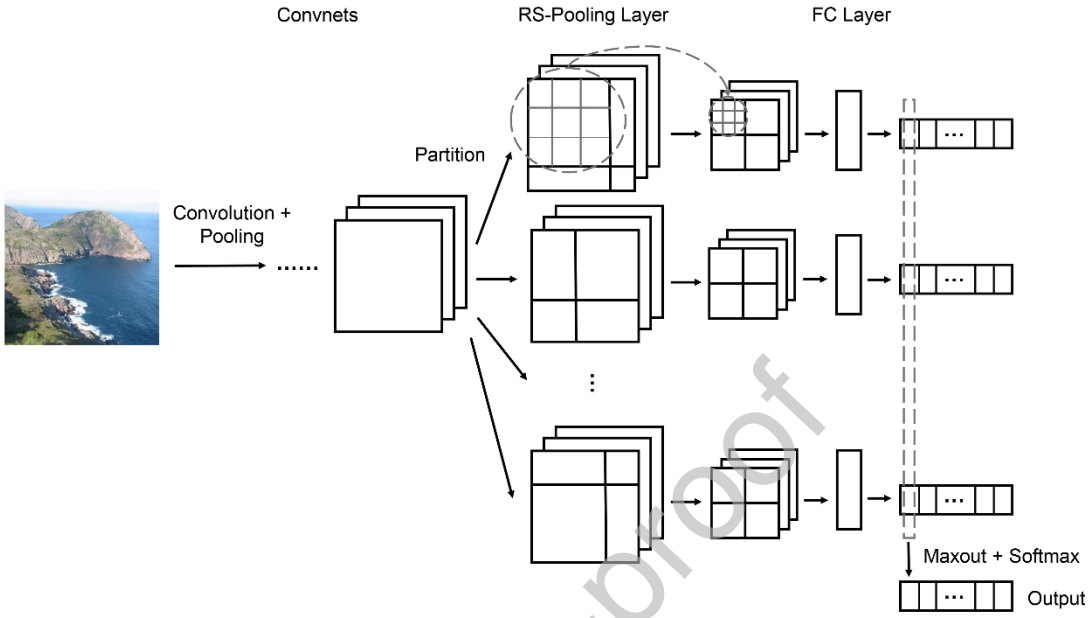

Fig. 9 Randomized spatial pooling layer is incorporated in CNNs (reproduced from [46]). The feature maps are partitioned into sub-regions, and then each sub-region is divided into 3×3 cells for pooling. The pooled features are fed to the FC layers. Finally the maxout and softmax layer are used to generate classification outputs.

3.4. Discriminative Region Detection

It is worth noting that the pooling operation in feature encoding and the partition scheme in spatial layout pattern learning may break the consistent regions into several fragments and result in the loss of some salient features. Fortunately, discriminative region detection can alleviate this problem and attempt to find the regions of interest. Object bank [48] is built on the response maps of a collection of object sensing filters pretrained on generic labeled objects, which integrates local semantic meanings into the whole image representation. When more objects are detected from scene images, the dimensionality of the response vector increases severely. Thus, a regularized logistic regression method is introduced to only activate dozens of instances for each class. The deformable part-based models [49] with latent SVM training to discover common visual structures can capture recurring visual elements and salient objects. ISPRs [53] is another

algorithm designed to learn region features and capture some subtle spatial location information at the same time. As presented in Fig. 10, a joint model for detecting discriminative regions and learning effective part appearance is built in ISPRs. For the same scene, most discriminative regions are located with a few clusters, which imply that the important parts have higher chances of occurring within these clusters, and the parts far away from the clusters do not contribute to the recognition of this class. Therefore, one goal of this model is to integrate these clusters to detect important regions and generate union features. Instead of extracting predefined local descriptors, the learnable root filter is used to convolve with the discriminative regions to extract local appearance. Therefore, the other goal of this model is to learn the corresponding root filter for the important regions of each class. The training process of this model is to alternately optimize the two problems. By the restriction of the learned locations and the specific root filters for each class, the false response caused by the influencing structure can be eliminated, which will make the pooling more robust to incorrect input.

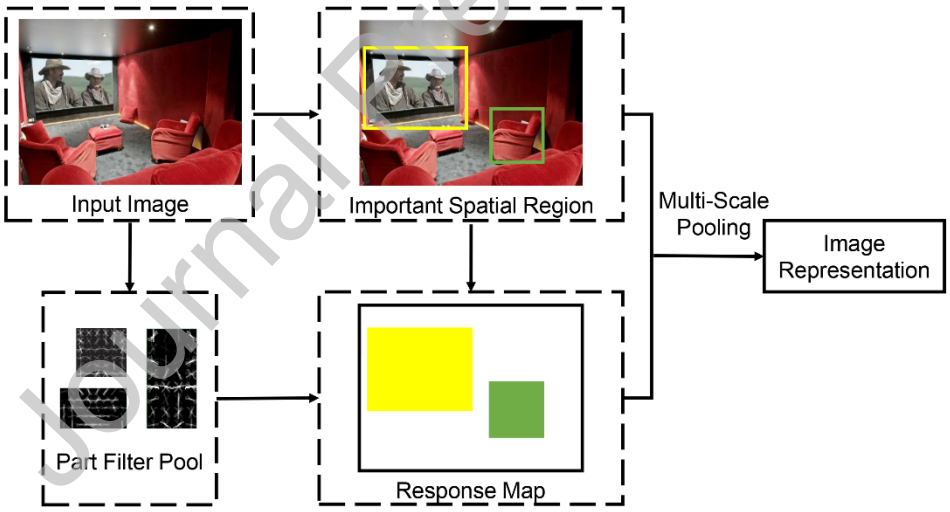

Fig. 10 The scheme of learning important spatial pooling regions (reproduced from [53]). The part response map is generated from the convolution of part filters and corresponding HOG maps. The important spatial pooling regions are introduced to process the response map with multi-scale pooling to form the image representation.

DSFL [54] is another interesting algorithm for region feature learning, in which a transformation filter bank is learned to transform pixel values of local image patches into features. To make the global filter bank more compact, only a subset

of filters is activated during the learning procedure. As shown in Fig. 11, some filters can be activated by many classes, which results in shareable features between different classes. The activation of filters is determined by a binary selection variable vector. Meanwhile, a discriminative term is introduced to force features from the same class to be close and those from different classes to be far away, which leads to discriminative features for each class. The discriminative features can be selected by nearest neighbor-based learning methods. In view of the expensive computation and potential noisy patches, a patch-to-database (P2D) distance is designed to measure the discriminative power of patch features. The entire feature learning is built on minimizing the error between the reconstructed data and the original data. Like deep architectures, DSFL can be stacked to extract multiple levels of features and produce better recognition performance. The ultimate image representation is constructed based on the LLC framework by utilizing the learned features.

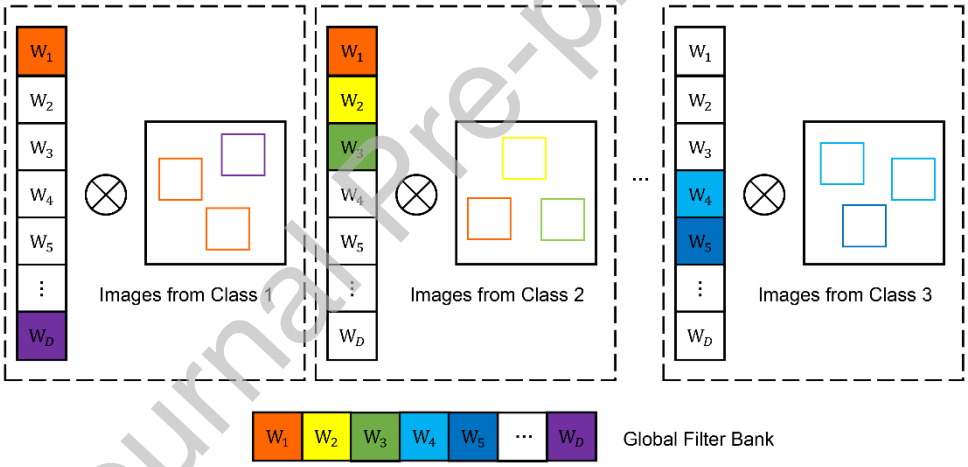

Fig. 11 Learning discriminative and shareable features for scene recognition (reproduced from [54]). $w_1, \ldots, w_D$ represent the filters in the global filter bank. A small subset of filters can be activated to learn the class-specific patterns for each class.

Discriminative regions can also be extracted in another relatively simple method. To imitate the principles of human visual attention, the visually sensitive region detection images can be obtained based on the context-based saliency detection algorithm [75]. This saliency detection images and the original images are superimposed to obtain the enhancement images in [55]. By feed the pre-trained CNNs with the original images, saliency detection images and enhancement images separately, the output deep features are concatenated to form the ultimate image representation for

scene recognition.

3.5. Object Correlation Analysis

The correlation of object configurations among different scenes has been exploited to distinguish various scenes. Some researchers attempted to build the co-occurrence pattern of some salient attributes by various probability analyses. These salient attributes are mostly hidden in some discriminative patches, so the extraction of these patches is very important. MetaObject-CNN [60] is proposed as a pipeline for harvesting discriminative objects for scene recognition. In this model, a region proposal technique is adopted to generate a set of salient patches potentially containing objects. For each scene category, the discriminative patch has a higher probability of appearing in the image and vice versa, so the class density estimation is conducted to implement patch screening. The convolutional features of the detected discriminative patches are aggregated into the image representation by VLAD. Similarly, inspired by the fact that the objects have different occurrence probabilities in various scenes, the recent work in [62] proposed to represent a scene by a bag of occurrence probabilities of discriminative objects. Specifically, the discriminative objects are selected from the posterior probabilities of scenes given objects, which are obtained from the object multinomial distributions in each scene using the Bayes rule. Semantic manifold models image patches as points in a semantic probability simplex; these patches are discovered by weak supervision via image labels, which involves the scene categories co-occurrence patterns. The neural network-based discriminative sematic multinomial and context models are proposed in [59] to improve the consistency of scene co-occurrence patterns.

3.6. Hybrid Deep Models

In the past few years, deep learning has dominated the field of computer vision due to its promising performance. Multiple levels of nonlinear operations allow the deep architecture to exploit the abstract information from the input images. Various architectures have been proposed for general image classification, including AlexNet [76], VGG [77], NiN [78], Inception [79], ResNet [80], DenseNet [81], and NASNet [82]. The outputs of CNNs retain abundant holistic features but

are short of local details. The DAG-CNN [63] combines the local features from lower layers and holistic features from top layers, which has achieved remarkable success. As shown in Fig. 12, DAG-CNN converts the chained feed-forward hierarchical architecture into a directed acyclic graph by connecting multiscale branches to the final classification layer. These connections introduce additional gradient flows from the output layer to the shallow layers, which relieves the vanishing gradients phenomena and helps CNNs to converge to a better local minimum for the classification. Other notable CNN architectures proposed at the same time and subsequent works, such as Inception, ResNet, and DenseNet, have also indicated similar effects. In general, features from intermediate and high layers (i.e., parts and objects) are more useful than those from low layers (i.e., edges and textures) in scene recognition. Features from nearby layers usually contain redundant or correlated information, which may reduce the recognition performance. A greedy forward-selection strategy is used to select appropriate features from different layers. Obviously, the performance advantages of the DAG-CNNs require exhaustive preliminary experiments.

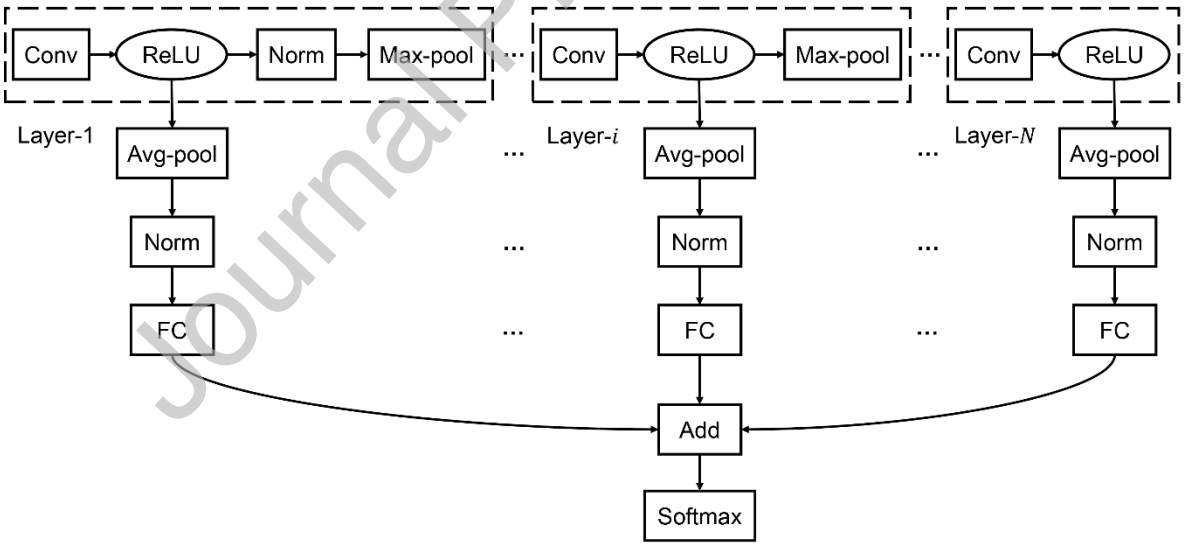

Fig. 12 The architecture of DAG-CNN (reproduced from [63]). Multiple branches are connected to an underlying chain backbone network. The outputs of the multiple branches are added to predict the class.

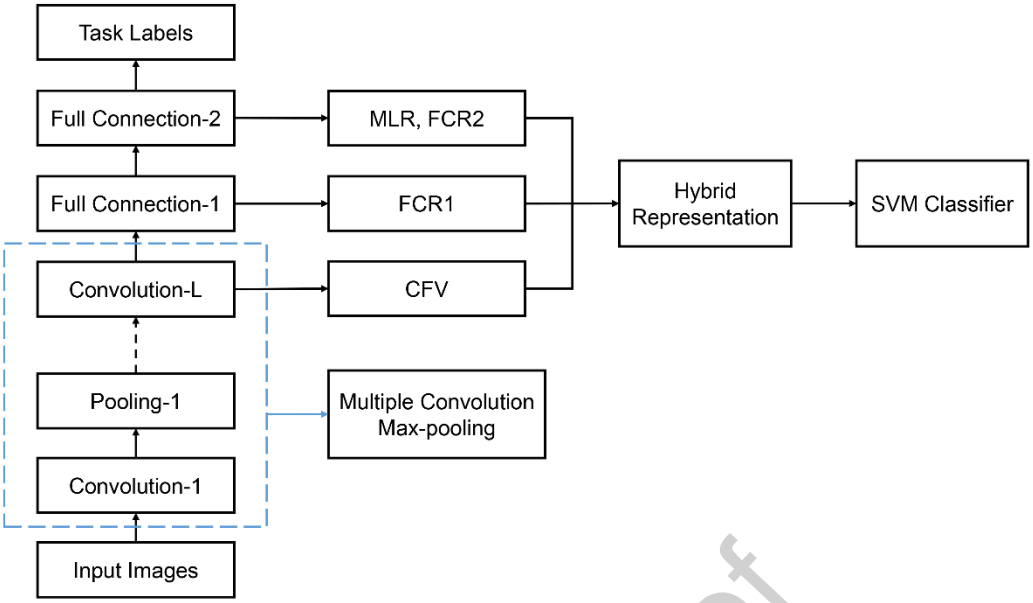

Fig. 13 Architecture of the hybrid model (reproduced from [65]). FCR1 and FCR2 indicate the global features extracted from the last two FC layers. CFV is calculated from the last convolutional feature, and MLR is calculated from the outputs of the last FC layer. Dashed box denotes the operations used in the CNN architecture.

Recently, many models that attempt to combine the strength of end-to-end networks and feature encoding have emerged. Feature encoding is conducted on patch features and global features to generate image representation, these features come from the multi-stage outputs of the end-to-end network. The hybrid model proposed in [65] combined CNNs with two feature encoding methods and led to better recognition accuracy. As shown in Fig. 13, one of the feature encoding method produces the mid-level local discriminative representation (MLR) using local patches. Firstly, local parts are extracted using selective search, then the two-stage clustering forms the codebook, finally LLC and SPM are conducted to produce the MLR. This work replaced the general one-stage codebook construction by the two-stage clustering process. Specifically, the location information of bounding boxes detected by the selective search and the corresponding last FC layer activations of these patches are used to construct the similarity graph for the spectral clustering. Some of the obtained clusters are selected to generate the class-specific and class-mixture codebooks by K-means clustering. The other feature encoding method is to attain the FVs of outputs of the last convolutional layer in CNNs (CFV), which corresponds to the global features. Both of them are combined to form the ultimate image representation for classification.

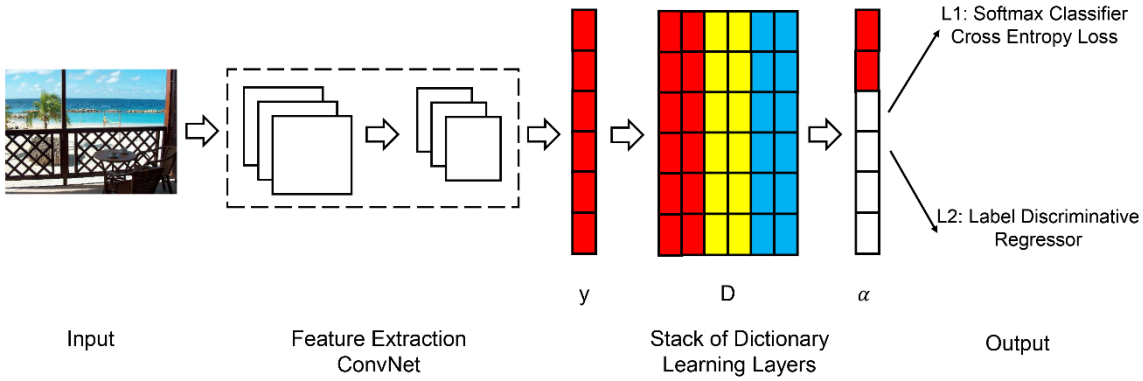

Fig. 14 Architecture of CNN-DL (reproduced from [67]). The FC layers are replaced by the dictionary learning layers. The L1 outputs the probabilities of all categories, and the L2 for the corresponding discriminant sparse representation depends on the learned dictionary.

To exploit the advantage of the sparse representation in CNNs, a new dictionary learning layer [67] is designed to automatically adapt different scene datasets with the proper sparse level. Instead of regarding the dictionary learning as separate post-processing on the CNNs features, the dictionary learning layer can be integrated into a unified deep network, as shown in Fig. 14. Learning all dictionary parameters is compatible with the backpropagation scheme in general CNNs, which helps to result in more adaptive sparse representations. Specifically, the dictionary learning layer is composed of several stacked recurrent units. There are two basic input variables for each recurrent unit: one is the sparse representation, and the other is the residual between convolutional features and reconstructed features based on the learnable dictionary and the sparse representation. The related parameters of the dictionary and convolutional features from previous layers can be learned by the backpropagation of gradients. In addition to the conventional cross-entropy loss function for fitting the prediction, a label discriminative regressor is proposed as a vital part of the final loss function. It is aimed at balancing the recognition accuracy and generalization ability by constraining the Mahalanobis distance on sparse representations and the Euclidean distance on the linear mapping matrix. The dictionary learning layer composed of a finite number of recurrent units is designed to mimic the alternate learning procedure of sparse coding and dictionary update, which exhibits a better convergence rate than Learned Iterative Shrinkage-Threshold Algorithm (LISTA) and decouples the shrinkage function threshold into two factors for better fitting of the unknown sparse prior. Apart from the sparsity control, this architecture

also emphasizes the discriminative capability of sparse codes.

In addition to the innovations of feature encoding based on CNNs, extra object datasets can be introduced to alleviate scene recognition. The mixed datasets of objects and scenes are constructed to train the Mix-CNN in [69]. To this end, the object labels and scene labels are combined for the training of Mix-CNN. The probabilities of all categories of local image patches are regarded as the Semantic multinomial (SMN). A codebook can be learned with the SMN extracted from the Mix-CNN. This codebook can be shared to encode the SMN extracted from Mix-CNN, Object-CNN and Scene-CNN. For the SMN extracted from Object-CNN and Scene-CNN, the corresponding part of the shared codebook is employed and the other part is discarded. The three kinds of encoded features can provide complementary representations to further improve the scene recognition performance.

Recently, FOSNet [70] is proposed to fuse the object and scene information in an end-to-end CNN framework without extra feature encoding methods, which includes Object-CNN stream and Scene-CNN stream. For the Scene-CNN stream, the scene coherence loss is proposed to impose the same category on adjacent patches of a single image. The adjacent patches correspond to the grid cells of feature maps in the CNN framework. The scene scores for each grid cell can be obtained by replacing the last global average pooling and FC layers with 1×1 convolution layers. For the Object-CNN stream, the class conversion matrix (CCM) is used to convert object features into the pseudo scene features. Finally, the new scene features are obtained by multiplying the original scene features from the Scene-CNN stream by the attention map generated from pseudo scene features from the Object-CNN stream in element-wise manner.

4. Available Datasets and Benchmarks

In this section, we introduce all publicly available datasets for scene recognition. Some of them have been widely used over the past few years, and some are new large-scale datasets to meet the requirements for the development of deep learning. Then, we will compare the recognition accuracy of some representative algorithms on these datasets and analyze the potential factors behind these results.

4.1 Public Available Datasets for Scene Recognition

(1) Scene-15 dataset: This is the early dataset collected for scene recognition [16]. It contains 4485 gray images of 15 different categories including natural and indoor scenes. The absence of color information may reduce the recognition accuracy for some algorithms based on CNNs. For each category, the number of images ranges from 200 to 400. The sizes of these images are relatively small. There are not separate training and test sets. In general, 100 training images for each category are randomly selected for training, and the remaining images are used for the test. To guarantee the reliability of the evaluation results, the random splits should be repeated many times.

(2) UIUC Sports-8 dataset: This dataset [83] contains 1572 color images in 8 different categories including various sports event scenes. The number of images for each category ranges from 130 to 250. These images have high resolutions (from 800×600 to thousands of pixels per dimension). It also does not provide separate training and testing sets. Following the protocol defined in [83], 70 images are randomly sampled for training and 60 images from the remaining images are randomly sampled for testing in each category. Multiple sampling is also necessary for reliable evaluation.

(3) MIT Indoor-67 dataset: It is a challenging indoor scene dataset, which contains 15620 color images including 67 categories. The indoor scenes have large intraclass variation, and there are many confusing indoor scenes due to the similar indoor backgrounds and some common objects. According to the standard partition of sets for evaluation in [84], the numbers of training images and testing images for each category are 80 and 20 respectively.

(4) SUN-397 dataset (Scene Understanding): This is a large-scale scene dataset. It includes 397 distinct scene categories and 108754 color images with at least 100 images per category. The categories comprise different kinds of indoor and outdoor scenes with tremendous objects and alignment variance, which brings more complexity for scene recognition. Following the standard protocol defined in [85], 100 images are selected from each category, where 50 images are used for training and the other 50 images are used for testing. The partitions are fixed and publicly available.

(5) Places dataset: Although the SUN-397 dataset provides a larger number of scene images, the quantity of each

category is still insufficient to feed deep learning models. Therefore, 2.5 million scene images are collected to form the Places dataset [86]. In Places-205, there are 205 common scene categories with at least 5000 images per category for training. For validation and testing, 100 and 200 images per category are used. Different from the average accuracy across all categories in the evaluation of those datasets mentioned above, the evaluation criteria of the Places are based on top-5 error. The Places2 dataset [87] is extended from the Places dataset, which contains more than 10 million images including more than 400 scene categories. This dataset is consistent with the real-world frequencies of occurrences and is probably the most challenging dataset for scene recognition currently. 365 categories with more than 4000 images each class are selected to create Places365-Standard and Places365-Challenge.

Table 1. Comparison of the public available scene datasets.

Dataset name	Number of scene categories	Number of total images	Number of images for training	Number of images for testing
Scene-15 [16]	15	4485	1500	2985
UIUC Sports-8 [83]	8	1572	560	480
MIT Indoor-67 [84]	67	15620	5360	1340
SUN-397 [85]	397	108754	19850	19850
Places-205 [86]	205	2.5M	2448873	41000
Places365-Standard [87]	365	10+M	1803460	328500
Places365-Challenge [87]	365	10+M	8M	328500

4.2 Comparison of Various Scene Recognition Algorithms

The average recognition accuracies of some representative scene recognition algorithms mentioned in this paper are presented in Table 2. Beyond the visible recognition accuracy, we will reveal some potential problems and delve more performance measurements including inference speed, training complexity and robustness. Except for the recognition accuracy, other measurements are hard to quantify, so we will give some empirical analysis.

Global attribute descriptors are obtained by some pre-defined numerical computations without the training process, so they have the fastest inference speed. However, they can only capture some simple visual features, which limit their expressive power and leads to the worst recognition accuracy. Meanwhile, these global attributes are susceptible to cluttered

backgrounds. For example, scenes with many buildings and cars in the suburbs may be recognized as city streets. Overall, global attribute descriptors are unstable and have the worst recognition accuracy, so they are unlikely to be applied in current scene recognition systems.

Patch feature encoding has been extensively explored in the past few years, and its recognition accuracy is relevant to the adopted patch features and codebook learning methods. As show in Table 2, deep features extracted from the patches using deep neural networks can achieve higher recognition accuracy than those handcrafted features. Deep neural networks can learn to capture the visual features most relevant to scene recognition, so it is obvious that deep features are more discriminative than pre-defined handcrafted features, but they require much more computational resources. Moreover, the advanced methods of codebook learning take more inherent relationships into account and lead to better results. For example, LR-Sc+SPM [22] can achieve higher recognition accuracy than LScSPM [20]. More advanced methods of codebook learning also result in more complicated training procedures. The algorithms based on patch feature encoding can deal with cluttered backgrounds and object deformation within a certain range, which are more robust than global attribute descriptors. However, a large number of sampled image patches greatly increase the inference time for deep neural networks. Enormous number of scene categories may overwhelm codebook learning and reduce recognition accuracy. As shown in Table 2, patch feature encoding equipped with CNNs can result in the highest recognition accuracy for the UIUC Sports-8 dataset, but the recognition accuracies for other datasets with more categories are relatively lower than other approaches. Overall, the algorithms based on patch feature encoding can be applied to some specific cases where computational resources and scene categories are limited while response time is more important than recognition accuracy.

The learning of spatial layouts patterns can provide more spatial structure information and improve the scene recognition accuracy, but there are upper bounds for this kind of algorithms because excessive spatial partitions may break some large scale objects into pieces, which has an adverse impact on the recognition accuracy. Some of them unify the spatial layouts pattern learning and classification training (e.g. RSP [43]), whereas some of them are designed as the extra

modules in the end-to-end customized networks (e.g. RS-Pooling [46], S2ICA [47]). Therefore, the training procedures of them are relatively simple and the inference speed is faster in comparison with most patch feature encoding based algorithms. Spatial layouts patterns learning is very effective for outdoor scenes and stable indoor scenes, but is also easily confused by highly similar indoor scenes with the same spatial layouts. Overall, this kind of algorithms can achieve the moderate recognition accuracy with minor modifications in current CNNs, which means they bring little extra computations compared with the existing architectures. They consume the least inference time against other approaches based on deep neural networks.

Discriminative region detection can provide complete regions of interest, and to some extent overcome the drawbacks resulted from excessive spatial partitions in spatial layouts patterns learning. Regions of interest can help cope with some complex scenes that are difficult for other kinds of algorithms. Therefore, this kind of algorithms can achieve better recognition results compared with those based on spatial layouts patterns learning as shown in Table 2. Some of them need pre-trained object detectors (e.g. Object Bank [48]), which makes the training procedure more complicated and consumes more time and computations. Some of them make the most of scene labels to autonomously select discriminative regions without object detectors (e.g. ISPRs [53]), but their inference time still depends on the region selection algorithms. In general, discriminative region detection is also sensitive to the number of scene categories because there are more kinds of objects for larger datasets with more categories. Overall, this kind of algorithms can achieve relatively higher recognition accuracy on small and intermediate datasets within a reasonable time period.

Object correlation analysis is most complex in all scene recognition approaches. The detection of the salient objects or discriminative patches is the prerequisite for the subsequent correlation analysis, which relies on the pre-trained object detector or some practical region proposal techniques. Various probability models are introduced to analyze the relationships between diverse objects and different scene categories. The increase of the number of object classes will lead to the curse of dimensionality for the feature vector that denotes the occurrence of all objects. Moreover, errors of detected

objects or patches will propagate to the subsequent correlation analysis. Therefore, the recognition accuracy not only relates to the model of object correlation, but also depends on the object detection model or the region proposal technique. As the result of abundant information, object correlation analysis can achieve moderate recognition accuracy, but the inference speed is the slowest because of the heavy computations. They are the optional choice only when object detection is required for the task.

Table 2. Recognition accuracy comparison of some representative algorithms on four public datasets.

Category	Approach	Feature Extraction	Scene-15	Sports-8	Indoor-67	SUN-397
Global Attribute Descriptors	GIST [1]	GIST	73.28	82.60	-	-
	CENTRIST [4]	CENTRIST	83.88	86.22	-	-
	LDBP [5]	LDBP	84.10	88.10	-	-
	mCENTRIST [6]	mCENTRIST	-	86.50	44.60	-
Patch Feature Encoding	SPM [16]	SIFT	81.40	81.80	34.40	-
	HIK [18]	CENTRIST	84.12	84.21	-	-
	LScSPM [20]	SIFT	89.75	85.31	-	-
	LR-Sc$^+$SPM [22]	SIFT	90.03	86.69	-	-
	DUCA [23]	AlexNet	94.50	98.70	71.80	-
	MOP-CNN [42]	AlexNet	-	-	68.88	51.98
	NNSD [88]	ResNet-152	94.70	**99.10**	85.40	64.78
Spatial Layouts Pattern Learning	RSP [43]	SIFT	88.10	79.60	-	-
	RS-Pooling [46]	AlexNet	89.40	-	62.00	-
	S^{2}ICA [47]	VGG-16	93.10	95.80	74.40	-
Discriminative Region Detection	Object Bank [48]	Object Filters	80.90	76.30	37.60	-
	ISPRs [53]	HOG	91.06	92.08	68.50	-
	DSFL [54]	AlexNet	92.81	96.78	76.23	-
	VS-CNN [55]	AlexNet	**97.65**	97.50	80.37	43.14
Object Correlation Analysis	MetaObject-CNN [60]	Hybrid CNN	-	-	78.90	58.11
	SDO [62]	VGG-16	95.88	-	86.76	73.41
Hybrid Deep Models	DAG-CNN [63]	VGG-19	92.90	-	77.50	56.20
	Hybrid CNNs [65]	VGG-19	-	-	82.24	64.53
	LS-DHM [63]	VGG-11	-	-	83.75	67.56
	Dual CNN-DL [67]	Hybrid CNN	96.03	-	86.43	70.13
	VSAD [68]	VGG-16	-	-	86.20	73.00
	Mix-CNN [69]	VGG-16	-	-	79.63	57.47
	FOSNet [70]	SE-ResNeXt-101	-	-	**90.37**	**77.28**
	Multi-scale CNNs [89]	VGG-16	95.18	-	86.04	70.17

Hybrid deep models benefit from elaborate architectures and extra representations or supervised information, which

can lead to the highest recognition accuracy on large-scale datasets. Some of them combine the expressive power of convolutional features and diverse feature encoding methods (e.g. Hybrid CNNs [65], CNN-DL [67], VSAD [68]), which results in complex training procedure, huge computational cost and longer inference time. In contrast, other customized end-to-end networks (e.g. DAG-CNN [63], Mix-CNN [69], FOSNet [70]) unify extra information into an architecture, which simplifies the training procedure. Meanwhile, they can obtain satisfied recognition accuracy and consume relatively less computations and inference time, but their consumed computations are still comparable or more than spatial layouts pattern learning and discriminative region detection due to their more complicated architectures. These customized end-to-end networks are favored with adequate computational resources.

5. Discussions

5.1 Effect of Multiple Scales

One naive multiscale approach is to regard CNNs as general feature extractors, which inevitably leads to dataset bias because the same fixed CNN models cannot adapt the data variation to different scales. For patch-based approaches, the distribution of object sizes with respect to the sampled patch varies over a wide range, and recognition performance gradually degrades as objects relative to scene categories decrease. For end-to-end networks, the density of objects per image between different scenes also greatly affects the receptive fields of CNNs; lower density causes CNNs to be better suited to object-centric datasets rather than scene datasets, which explains why CNNs trained on ImageNet are limited when used for scene recognition. Many works have been built on existing CNN models pre-trained on large-scale image datasets, such as ImageNet and Places. The selection of suitable patch scales for ImageNet-CNN and Places-CNN is noted in [89]. Their extensive experiments demonstrated that ImageNet-CNN exhibited good performance on patches at intermediate scales and that Places-CNN achieved the best performance on patches at the global scene level. Naturally, it makes sense to combine these two complementary deep features to boost the recognition performance. The straight practice is to splice the deep features from two scale-specific networks: one is Places-CNN extracting features at global scales, and

the other is ImageNet-CNN extracting features from patches at more local scales. The deep features here are the activation output of the last FC layer. Extracting patches from images resized to different sizes is an alternative to consider the multiscale factor. Object scale patches are fed to the ImageNet-CNN, while scene scale patches are fed to the Places-CNN. The two resulting activations from the last FC layer are concatenated into the image representation to train the SVM. For the object scale and scene scale, they are selected based on the evaluation of available datasets using pairwise combinations at a range of different scales. Their further experimental results suggested that more scales would be marginally helpful, so this dual architecture seems to be a compromising scheme to decrease unnecessarily heavy costs. Even so, the selection of scales is still an exhausting task when we build a suitable architecture for new target datasets.

5.2 Codebook Learning in Patch Feature Encoding

For patch feature encoding, the construction of the codebook has critical impact on recognition performance. On one hand, the potential defect of codebook learning is that superior performance depends on the appropriate dictionary size (the number of visual words) for each category. Codebook learning seems to be expensive and unresolvable when it comes to enormous amount of scene images. In addition, the dimensionality of the derived codes drastically increases when the number of categories becomes very large, which will lead to more complexity and slower inference processes. To address this problem, it is necessary to learn a compact codebook while simultaneously maintaining high recognition performance. By introducing an indicator function to remove correlated visual words, the Automatic Compact Dictionary Learning (ACDL) [90] method is proposed to decrease the codebook size. On the other hand, some constraints are employed to guarantee the specific characteristic of the derived codes, such as sparsity, discrimination, and selectivity. The comprehensive analysis presented in [91] reveals the main characteristics of various codebook learning methods. However, the selection of appropriate characteristics and the balance of different characteristics require further exploration.

5.3 Issues on End-to-end Networks

Although there have been some large-scale scene datasets for training, current end-to-end networks cannot attain the

best recognition accuracy as they are in the field of object recognition. This might be attributed to the three obstacles: ambiguous categorization, complicated spatial layouts and diverse contexts. Due to the uncertainty of scene concepts and existing overlap among different categories, label ambiguity is a challenge for the training of end-to-end networks. To relieve the training difficulty, some extra knowledge [92] is exploited to generate new labels to guide networks to a better optimization and reduce the effect of overfitting. The extra knowledge includes the correlation of different categories computed from the confusion matrix and knowledge networks pre-trained on relatively smaller and well-labeled datasets. Although some customized modules have been proposed to remedy the lack of spatial layout information as mentioned in section 3.3, all of them are designed by means of the predefined partition schemes. It is necessary to explore more flexible and learnable spatial partition modules in consideration of the complicated spatial layouts in scene images. Diverse backgrounds and various relations between foreground objects impede the representation learning in deep models. To introduce more contextual information, hybrid deep models have shown considerable advantages. However, some of these models also suffer from the same problems encountered in codebook learning. At the same time, the combination of codebook learning and deep models also consumes more memory and time. Integrating existing codebook learning methods into deep networks may be a promising alternative, which is still an open problem.

In addition, the long-tailed distribution of scene images is often encountered. The long-tailed distribution means that there are a few common classes (head classes) and many rare classes (tail classes) in our real world. To overcome this problem, dynamic meta-embedding module [93] is proposed to learn the transferable knowledge between head and tail classes, whereas the modulated attention module is employed to discriminate features of different classes. On the one hand, the dynamic meta-embedding combines the original deep feature and the memory feature by the learnable concept selector (a lightweight network). The learning of the memory feature adopts the discriminative centroids of the deep features of different classes as the basic building block, in which both intra-class compactness and inter-class discrimination are considered. On the other hand, since the location distributions of key cues of head and tail classes seem to be different, the

conditional spatial attention is applied to the self-attention map to help maintain the discrimination between head and tail classes.

5.4 Future Directions

After comparing various state-of-the-art approaches in scene recognition described above, we will then give some future directions, which may be helpful for further research in this field. To attain high accuracy for scene recognition, local information and global appearance are both indispensable. Global appearance can be considered as the complementary features for patch feature encoding. Extracting patch features on response maps instead of image patches provides a promising alternative to improve the efficiency and reduce the computational complexity. Recently, the attention mechanism is also applied to image recognition tasks [94], which can be explored to relieve the discriminative region detection and retain global features. Spatial layout pattern learning can be designed as a pluggable module to provide extra spatial structure information for object correlation analysis and hybrid deep models. The outputs of Object-CNN and Scene-CNN are combined in some hybrid deep models, which may be suboptimal for scene recognition. Extra supervised information (e.g. soft labels of objects in [92]) can be introduced to design a multi-task learning deep model that exploits local objects features and global appearance simultaneously.

6. Conclusion

In this paper, we have reviewed various scene recognition algorithms in terms of the feature transformation. All of them can be grouped into six major categories: global attribute descriptors, patch feature encoding, spatial layout pattern learning, discriminative region detection, object correlation analysis and hybrid deep models. Some representative works have been introduced in detail for each category, and their motivation and improvement are revealed. We also compare different categories of algorithms from the aspect of recognition performance and potential problems. Several issues are discussed for future research. Although substantial experiments have shown that multiscale ensemble can lead to drastic improvement in recognition performance, the appropriate selection of features at multiple scales is still task-dependent.

Meanwhile, multiscale ensemble consumes more computational resources and slows down the process. Instead of using multiple deep models or multiscale inputs, it is necessary to devise an efficient deep model. Feature encoding has shown new vitality in combination with CNNs. Hybrid deep models have already become the promising trend in scene recognition.

Conflict of interest

None declared.

Acknowledgements

This research is partially supported by The Programme for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, and also partially supported by JSPS KAKENHI Grant Number 15K00159.

References

[1] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelop, International Journal of Computer Vision, 42 (3) (2001) 145-175.

[2] J. Vogel, B. Schiele, A semantic typicality measure for natural scene categorization, in: Joint Pattern Recognition Symposium, 2004, pp. 195-203.

[3] A. Payne, S. Singh, Indoor vs outdoor scene classification in digital photographs, Pattern Recognition, 38 (2005) 1533-1545.

[4] J. Wu, J.M. Rehg, Centrist: a visual descriptor for scene categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1489-1501.

[5] X. Meng, Z. Wang, L. Wu, Building global image features for scene recognition, Pattern Recognition 45 (2012) 373-380.

[6] Y. Xiao, J. Wu, J. Yuan, mCENTRIST : a multi-channel feature generation mechanism for scene categorization, IEEE Transactions on Image Processing 23 (2) (2014) 823-836.

[7] T. Ojala, M. Petikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recognition 29 (1996) 51-59.

[8] D.G. Lowe, Distinctive image features from scale-invariant key-points, International Journal of Computer Vision 60 (2) (2004) 91-110.

[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[10] H. Bay, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 404-417.

[11] R. Margolin, L. Zelnik-Manor, A. Tal, OTC: a novel local descriptor for scene classification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 377-391.

[12] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 1470-1477.

[13] G. Csurka, C. Bray, C. Dance, L. Fan, Visual categorization with bags of keypoints, in: Proceedings of the European Conference on Computer Vision, 2004, pp. 1-22.

[14] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 524-531.

[15] K. Graumanand, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 1458-1465.

[16] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169-2178.

[17] J.C. van Gemert, J. Geusebroek, C.J. Veenman, A.W.M. Smeulder, Kernel codebooks for scene categorization, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 696-709.

[18] J. Wu, J.M. Rehg, Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 630-637.

[19] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794-1801.

[20] S. Gao, I.W.H Tsang, L.T. Chia, P. Zhao, Local features are not lonely – Laplacian sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3555-3561.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Locality-constrained linear coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360-3367.

[22] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, S. Ma, Image classification by non-negative sparse coding, low-rank and sparse decomposition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1673-1680.

[23] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, F.A. Sohel, A discriminative representation of convolutional features for indoor scene recognition, IEEE Transactions on Image Processing, 25 (7) (2016) 3372-3383.

[24] J. Qin, N.H.C Yung, Scene categorization via contextual visual words, Pattern Recognition, 43 (5) (2010) 1874-1888.

[25] N. M. Elfiky, F. S. Khan, J. Weijer, J. Gonzalez, Discriminative compact pyramids for object and scene recognition, Pattern Recognition, 45 (2012) 1627-1636.

[26] L. Zhou, Z. Zhou, D. Hu, Scene classification using a multi-resolution bag-of-features model, Pattern Recognition, 46 (1) (2013) 424-433.

[27] M. Clement, C. Kurtz, L. Wendling, Learning spatial relations and shapes for structural object description and scene recognition, Pattern Recognition, 84 (2018) 197-210.

[28] V. Sydorov, M. Sakurada, C. H. Lampert, Deep Fisher kernels -- end to end learning of the Fisher kernel GMM parameters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1402-1409.

[29] Y. Yuan, L. Mou, X. Lu, Scene recognition by manifold regularized deep learning architecture, IEEE Transactions on Neural

Networks and Learning Systems 26 (10) (2015) 2222-2233.

[30] H. Goh, N. Thome, M. Cord, J.H. Lim, Learning deep hierarchical visual feature coding, IEEE Transactions on Neural Networks and Learning Systems, 25 (2014) 2212-2225.

[31] G. Xie, X. Zhang, C. Liu, Efficient feature coding based on auto-encoder network for image classification, in: Proceedings of the Asian Conference on Computer Vision, 2014, pp. 628-642.

[32] L. Xie, F. Lee, L. Liu, Z. Yin, Y. Yan, W. Wang, J. Zhao, Q. Chen, Improved spatial pyramid matching for scene recognition, Pattern Recognition, 82 (2018) 118-129.

[33] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulder, J. Geusebroek, Visual word ambiguity, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7) (2009) 1271-1283.

[34] Y. Boureau, F. Bach, Learning mid-level features for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2559-2566.

[35] X. Zhou, X. Zhuang, H. Tang, M.H. Johnson, T.S. Huang, Novel gaussianized vector representation for improved natural scene categorization, Pattern Recognition Letter 31 (8) (2010) 702-708.

[36] T. Harada, Y. Ushiku, Y. Yamashita, Y. Kuniyoshi, Discriminative spatial pyramid, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1617-1624.

[37] L. Xie, J. Wang, B. Guo, B. Zhang, Q. Tian, Orientational pyramid matching for recognizing indoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3734-3741.

[38] T. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: Proceedings of the Advances in Neural Information Processing Systems, 1998.

[39] J. Sanchez, F. Perronnin, T. Mensink, J.J. Verbeek, Image classification with the Fisher vector: theory and practice, International Journal of Computer Vision, 105 (3) (2013) 222-245.

[40] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE

Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1704-1716.

[41] M. Cimpoi, S. Maji, A. Vedaldi, Deep filter banks for texture recognition and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3828-3836.

[42] Y. Gong, L. Wang, R. Guo, Multi-scale orderless pooling of deep convolutional activation features, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 392-407.

[43] Y. Jiang, J. Yuan, G. Yu, Randomized spatial partition for scene recognition, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 730-743.

[44] C. Weng, H. Wang, J. Yuan, X. Jiang, Discovering class-specific spatial layouts for scene recognition, IEEE Signal Processing Letters, 24 (8) (2017) 1143-1147.

[45] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2014) 346-361.

[46] M. Yang, B. Li, H. Fan, Y. Jiang, Randomized spatial pooling in deep convolutional networks for scene recognition, in: Proceedings of the IEEE Conference on Image Processing, 2015, pp. 402-406.

[47] M. Hayat, S.H. Khan, M. Bennamoun, S. An, A spatial layout and scale invariant feature representation for indoor scene classification, IEEE Transactions on Image Processing, 25 (10) (2016) 4829-4841.

[48] L. Li, H. Su, L. Fei-Fei, E.P. Xing, Object bank: a high-level image representation for scene classification and semantic feature sparsification, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 1378-1386.

[49] M. Pandey, S. Lazebnik, Scene recognition and weakly supervised object localization with deformable part-based models, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1307-1314.

[50] S. Singh, A. Gupta, A. A. Efros, Unsupervised discovery of midlevel discriminative patches, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 73-86.

[51] M. Juneja, A. Vedaldi, C. V. Jawahar, A. Zisserman, Blocks that shout: distinctive parts for scene classification, in: Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 923-930.

[52] Y. Yuan, J. Wan, Q. Wang, Congested scene classification via efficient unsupervised feature learning and density estimation, Pattern Recognition, 56 (2016) 159-169.

[53] D. Lin, C. Lu, R. Liao, J. Jia, Learning important spatial pooling regions for scene regions for scene classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3726-3733.

[54] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, X. Jiang, Learning discriminative and shareable features for scene classification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 552-568.

[55] J. Shi, H. Zhu, S. Yu, W. Wu, H. Shi, Scene categorization model using deep visually sensitive features 7 (2019) 45230-45239.

[56] L. Cao, L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1-8.

[57] Z. Niu, G. Hua, X. Gao, Context aware topic model for scene recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2743–2750.

[58] S.N. Parizi, J.G. Oberlin, P.F. Felzenszwalb, Reconfigurable models for scene recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2775–2782.

[59] X. Song, S. Jiang, L. Herranz, Multi-scale multi-feature context modeling for scene recognition in the semantic manifold, IEEE Transactions on Image Processing, 26 (6) (2017) 2721-2735.

[60] R. Wu, B. Wang, W. Wang, Y. Yu, Harvesting discriminative meta objects with deep CNN features for scene classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1287-1295.

[61] X. Song S. Jiang L. Herranz Y. Kong and K. Zheng Category Co-occurrence Modeling for Large Scale Scene Recognition, Pattern Recognition 59 (2016) 98-111.

[62] X. Cheng, J. Lu, J. Feng, B. Yuan, J. Zhou, Scene recognition with objectness, Pattern Recognition, 74 (2018) 474-487.

[63] S. Yang, D. Ramanan, Multi-scale recognition with DAG-CNNs, in: Proceedings of the IEEE International Conference on Computer

Vision, 2015, pp. 1215-1223.

[64] P. Tang, H. Wang, S. Kwong, G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition, Neurocomputing, 255 (2017) 188-197.

[65] G. Xie, X. Zhang, S. Yan, C. Liu, Hybrid CNN and dictionary-based models for scene recognition and domain adaptation, IEEE Transactions on Circuits and Systems for Video Technology, 27 (6) (2017) 1263-1274.

[66] S. Guo, W. Huang, L. Wang, Y. Qiao, Locally supervised deep hybrid model for scene recognition, IEEE Transactions on Image Processing, 26 (2) (2017) 808-820.

[67] Y. Liu, Q. Chen, W. Chen, I. Wassell, Dictionary learning inspired deep network for scene recognition, in: Proceedings of AAAI Conference on Artificial Intelligence, 2018, pp. 7178-7185.

[68] Z. Wang, L. Wang, Y. Wang, B. Zhang, Y. Qiao, Weakly supervised PatchNets: describing and aggregating local patches for scene recognition, IEEE Transactions on Image Processing, 26 (4) (2017) 2028-2041.

[69] S. Jiang, G. Chen, X. Song, L. Liu, Deep patch representations with shared codebook for scene classification, ACM Transactions on Multimedia Computing, Communications, and Applications 15 (1) (2019).

[70] H. Seong, J. Hyun, E. Kim, FOSNet: an end-to-end trainable deep neural network for scene recognition, arXiv abs/1907.07570.

[71] X. Wei, S. L. Phung, A. Bouzerdoum, Visual descriptors for scene categorization experimental evaluation, Artificial Intelligence Review 45 (3) (2016) 333-368.

[72] E. Anu, K.S. Anu, A Survey on Scene Recognition, International Journal of Science, Engineering and Technology Research (IJSETR) 5 (1) (2016) 64-68.

[73] V. Singh, D. Girish, A. Ralescu, Image understanding - a brief review of scene classification and recognition, in: Proceedings of Modern Artificial Intelligence and Cognitive Science (MAICS), 2017, pp. 85-91.

[74] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[75] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (10) (2012) 1915-1926.

[76] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

[77] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015.

[78] M. Lin, Q. Chen, S. Yan, Network in network, in: Proceedings of the International Conference on Learning Representations, 2014.

[79] L. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv abs/1706.05587.

[80] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[81] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261-2269.

[82] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697-8710.

[83] L. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1-8.

[84] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 413-420.

[85] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, A. Torralba, SUN database: large-scale scene recognition from abbey to zoo, in: Proceedings of the IEEE International Conference on Computer Vision, 2010, pp. 3485-3492.

[86] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using Places database, in:

Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 487-495.

[87] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2018) 1452-1464.

[88] L. Xie, F. Lee, L. Liu, Z. Yin, Q. Chen, Hierarchical Coding of Convolutional Features for Scene Recognition, IEEE Transactions on Multimedia, (2019, Early Access). https://doi.org/10.1109/TMM.2019.2942478.

[89] L. Herranz, S. Jiang, X. Li, Scene recognition with CNNs: objects, scales and dataset bias, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 571-579.

[90] Y. Song, Z. Zhang, L. Liu, A. Rahimpour, H. Qi, Dictionary reduction: automatic compact dictionary learning for classification, in: Proceedings of the Asian Conference on Computer Vision, 2017, pp. 305-320.

[91] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature coding in image classification: a comprehensive study, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 493-506.

[92] L. Wang, S. Guo, W. Huang, Y. Xiong, Y, Qiao, Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs, IEEE Transactions on Image Processing, 26 (4) (2017) 2055-2068.

[93] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, S. X. Yu, Large-scale long-tailed recognition in an open world, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[94] L. Drew, S. Dan, E. Sven, S. Thomas, Learning what and where to attend, in: Proceedings of the International Conference on Learning Representations, 2019.

AUTHOR BIOGRAPHY

Lin Xie received his M.S. degree in control science and engineering from University of Shanghai for Science and Technology. His current research interests are in the areas of scene recognition and machine learning.

Feifei Lee received her Ph.D. degree in electronic engineering from Tohoku University in Japan, in 2007. She is currently a professor at the University of Shanghai for Science and Technology. Her research interests include pattern recognition, video indexing, and image processing.

Li Liu received the Ph.D. degree in pattern recognition and intelligent system from East China Normal University, Shanghai, China, in 2015. She was with the Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, QC, Canada, from 2013 to 2014 as a visiting doctoral student, and in 2016 as a visiting scholar. She is currently a lecturer with Nanchang University. Her research interests include pattern recognition, computer vision, and document image analysis.

Koji Kotani received the B.S., M.S. and Ph.D. degrees all in electronic engineering from Tohoku University, Japan, in 1988, 1990 and 1993, respectively. He is currently a professor at Department of Electronics and Information Systems, Akita Prefectural University. He is engaged in the research and development of high performance devices/circuits as well as intelligent electronic systems. Dr. K. Kotani is a member of IEEE and a member of the Institute of Electronics, Information and Communication Engineers of Japan.

Qiu Chen received Ph.D. degree in electronic engineering from Tohoku University, Japan, in 2004. Since then, he has been an assistant professor and an associate professor at Tohoku University. He is currently a professor at Kogakuin University. His research interests include pattern recognition, computer vision, information retrieval and their applications. He is also a guest professor at the

University of Shanghai for Science and Technology. Dr. Chen serves on the editorial boards of several journals, as well as committees for a number of international conferences

Declaration of interests

☑The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: