Chapter 2
Scene Understanding Datasets

Keywords Dataset · Large-scale · PASCAL dataset · ImageNet · LabelMe dataset · Fifteen scene category dataset · CMU 300 dataset · Tiny image dataset · SUN dataset · PLACE205 dataset

2.1 Small-Scale Scene Understanding Datasets

At early ages of scene understanding, several benchmarks were proposed for research purposes. The 8-scene dataset, the 15-scene dataset, the UIUC sport dataset were dominate ones. These traditional datasets have several common properties. First, they consist of typical and iconic scene images that background objects and surfaces are the determinative components to decide scene categories. Besides, these datasets have limited number of scene categories and images within the same class have little variations in visual patterns. Finally, they focus more on usually-seen categories, such as "coast" and "living room", etc., and do not have seldom-seen scenes or detailed categories such as "temple", "auditorium", etc.

2.1.1 8-scene Dataset

8-scene dataset was firstly used in the work [1] where the famous image descriptor "GIST" was proposed. It consists of 8 outdoor scene categories (4 of them are from natural landscapes and 4 of them are from man-made scenes). Totally, there are 2688 images in the dataset. It has been treated as the most famous benchmark in the scene understanding field and a mandatory challenge for all scene understanding researches since then. Later, researchers from the same group labeled the dataset for

C. Chen et al., *Big Visual Data Analysis*,
SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-0631-9_2

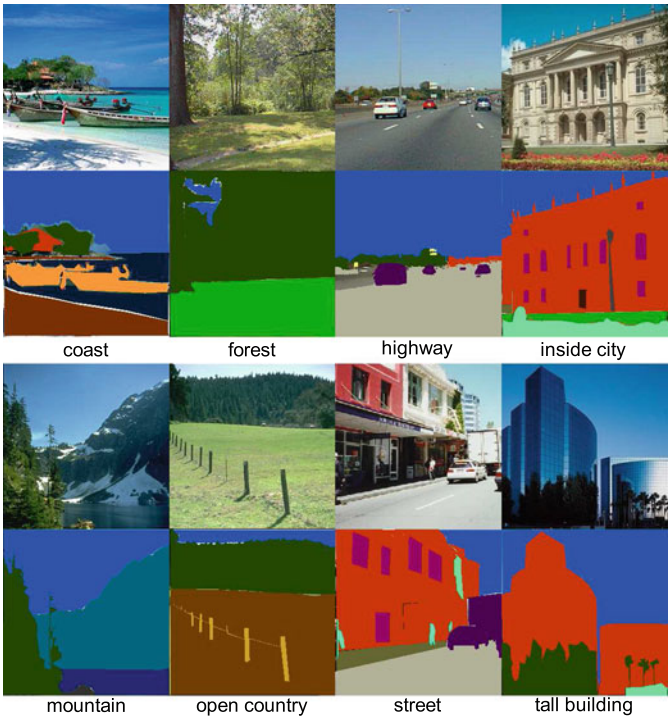

Fig. 2.1 Examples in 8-scene dataset

the purposes of semantic segmentation in scene parsing researches. The segmentally labeled data was usually called SIFTFlow after the work [2]. Examples images from the 8 scenes and corresponding segmental labels can be seen in Fig. 2.1.

2.1.2 15-scene Dataset

After the blooming of 8-scene [1], researchers intended to increase the categories and diversities to create more challenging datasets. Inheriting the popularity of the 8-scene dataset, 15-scene dataset includes 2 extra outdoor categories and 5 indoor categories to meet the requirement on general scene understanding tasks. As shown in Fig. 2.2, proposed by [3, 4], the 15-scene dataset has three more challenging factors than the 8-scene datasets: (1) the 15-scene dataset consists of only gray-scale images, (2) the 15-scene dataset has twice number of scene categories and images and (3) the 15-scene dataset considers indoor scenes with outdoor scenes together. With these challenges and differences, bonded with the 8-scene dataset, the 15-scene dataset also became a mandatory benchmark in nowadays researches.

Fig. 2.2 Examples in 15-scene dataset: stared categories are originally from the 8-scene dataset

2.1.3 UIUC Sports

UIUC sports [5] contains 8 sports event categories: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). Images are divided into easy and medium according to the human subject judgment. Information of the distance of the foreground objects is also provided for each image. Unlike the 8-scene dataset and the 15-scene dataset, UIUC sports is an event-centric scene dataset. As shown in Fig. 2.3, most of its scene images have distinctive foreground and background contexts. Therefore it arouses a promising research direction in context based recognition using scene/object topic models.

2.1.4 CMU 300

The dataset in [6] is the largest benchmarking dataset in the geometric layout research community. It consists of 300 images of outdoor scenes with 23 different scene categories including alley, building, cliff, college etc. Example images are shown in Fig. 2.4. The dataset provides ground truth geometric labels for each image, namely support, sky, planar left, planar center, planar right, non-planar solid and porous. The first 50 images are used for training the surface segmentation algorithm as done in previous work [6, 7]. The remaining 250 images are used for evaluation. Besides geometric labels, it also provides occlusion boundaries for 100 images in this dataset. The occlusion boundaries indicate the depth orders of occluding and occluded objects.

Fig. 2.3 Examples in UIUC sports dataset

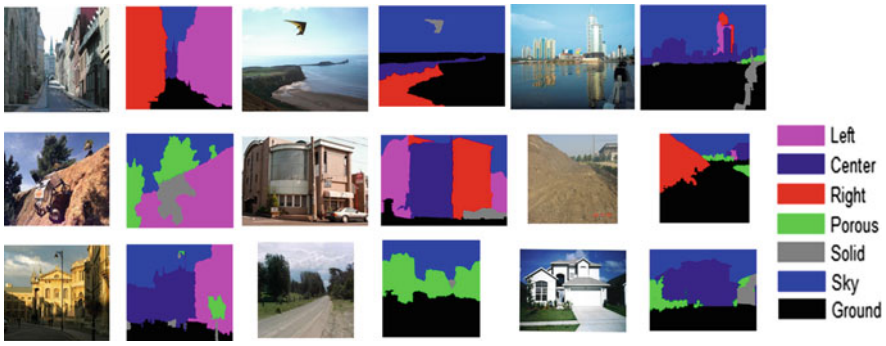

Fig. 2.4 Examples in CMU 300 dataset

2.2 Large-Scale Scene Understanding Datasets

With progresses made in computer vision research, small traditional image datasets are no longer sufficient for the performance evaluation purpose of robust scene understanding system, and several large scale datasets are built to meet the urgent need. We will present several of them commonly used for scene understanding.

2.2.1 80 Million Tiny Image Dataset

The 80 million tiny image dataset contains 7,527,697 images. It was built by [8]. Each image is of low resolution (with image size dimension 32×32) and labeled with one of the 53,464 English nouns from the WordNet [9]. It is mainly used in fast image search which demands very little memory. All images were obtained from the Google search and other engines using English nouns from the WordNet. These images are with high diversity, containing object images and scene images.

The construction of this dataset was motivated by an interesting experimental observation. That is, human can classify a scene with 32×32 pixels and achieve a recognition rate higher than 80% [8]. The purpose of creating this dataset is to facilitate the development of fast image search and scene matching techniques with very little memory. It combines Vogel and Schiele [10] 702 natural scenes, Olivia and Torralba's [1] 2688 images, Caltech 101 categories, Caltech 256 categories. The Caltech 101 categories and Caltech 256 categories are images containing objects widely used in object recognition. The 80 million tiny image dataset contains image categories of sufficient diversity. Although the image number is large, all images are categorized.

In the project website [8], it provides the confidence map, labels provided by users, nouns from the WordNet and the visual dictionary view of the whole dataset. The confidence map reflects the algorithmic accuracy in the classification of different categories. The labels show the amount of data provided by users. It also offers the mosaic image after images are divided into semantic groups. An example can be seen in Fig. 2.5. The visual dictionary view provides the visualization of tiles by averaging the color of images that have the same English noun. The averaged version of a class of images can reflect the global information of this class. The website allows users to add annotations by selecting a word and indicating whether correct and wrong

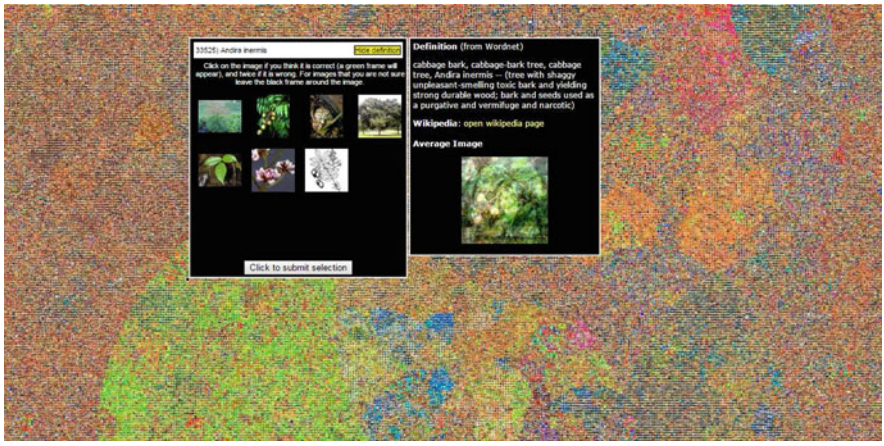

Fig. 2.5 The poster of the visual dictionary built in the tiny image dataset

images are returned. To fix inaccurate annotations, it allow users to correct labels online. However, the low resolution of these tiny images usually do not allow image classification algorithms to work properly.

2.2.2 *PASCAL Dataset*

The PASCAL Visual Object Classes (VOC) Challenge dataset [11] provides standardized image datasets for object class recognition and a common set of tools to access the data sets and annotations. There are 20 object classes in the PASCAL dataset with thousands of images in each class. Examples of object images can be seen in Fig. 2.6. The images were obtained from the collection of Flicker photos. The PASCAL VOC Challenge was an annual event from 2005 to 2012, in which researchers submitted their results on object classification, and got their results evaluated and compared online. The competition includes:

- Classification: determining presence/absence of an example of that class in the test image for each of the 20 classes.
- Detection: determining the bounding box and the label of each object from the 20 target classes in the test image.
- Segmentation: generating the pixel-wise segmentation of an object in the image.
- Person layout: determining the bounding box and the label of each part of a person.
- Action classification: determining the action of a person in a still image.

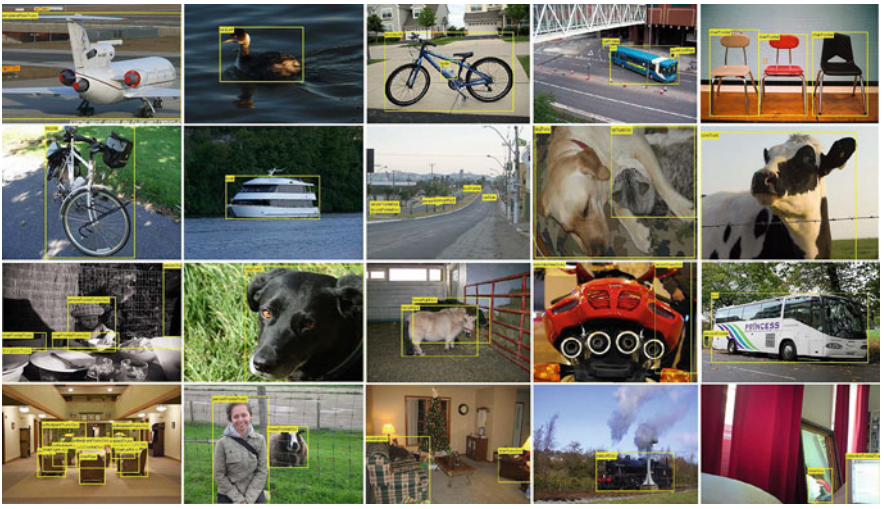

Fig. 2.6 The 20 object classes in PASCAL dataset

Although the PASCAL dataset is a large and challenging dataset for object classification and recognition, it is not an appropriate dataset for our interest in dealing with scene images.

2.2.3 ImageNet Dataset

The ImageNet [12] is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in the WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, and a great majority of them are nouns (80,000+). The ImageNet aims to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. Compared to the other image classification dataset, the ImageNet is the largest and most challenging dataset for object classification and recognition. On the other hand, it focuses on general image classification challenges, which include scene classification as only a small branch of the problem. Figure 2.7 shows several examples in the dataset. We can see most of them are foreground objects oriented.

2.2.4 LabelMe Dataset

LabelMe [13] is a project conducted by the MIT CSAIL with an objective to provide a dataset of digital images with object and surfaces annotations. The dataset is dynamic, free to use, and open to public contribution. Specifically, it provides a website for people to annotate images online. An example can be seen in Fig. 2.8. LabelMe asks users to use polygons to segment and annotate object and surfaces in an image. As of October 31, 2010, LabelMe has 187,240 images, 62,197 annotated images, and 658,992 labeled objects. The project was motivated by the following observation. Most available data in computer vision research are tailored to the problem of a specific research group and it is often that new researchers need to collect additional data to solve their own problems. LabelMe was created to solve this shortcoming. LabelMe are different from other existing datasets in the following aspects. First, LabelMe contains images of objects with multiple angles, sizes and orientations. Second, LabelMe designs images for object recognition in arbitrary scenes and it avoids the scene to cropped, normalized or resized. Third, each image in LabelMe may contain more than one object, and users are allowed to label these objects. Finally, its numbers of images and object classes can be easily increased.

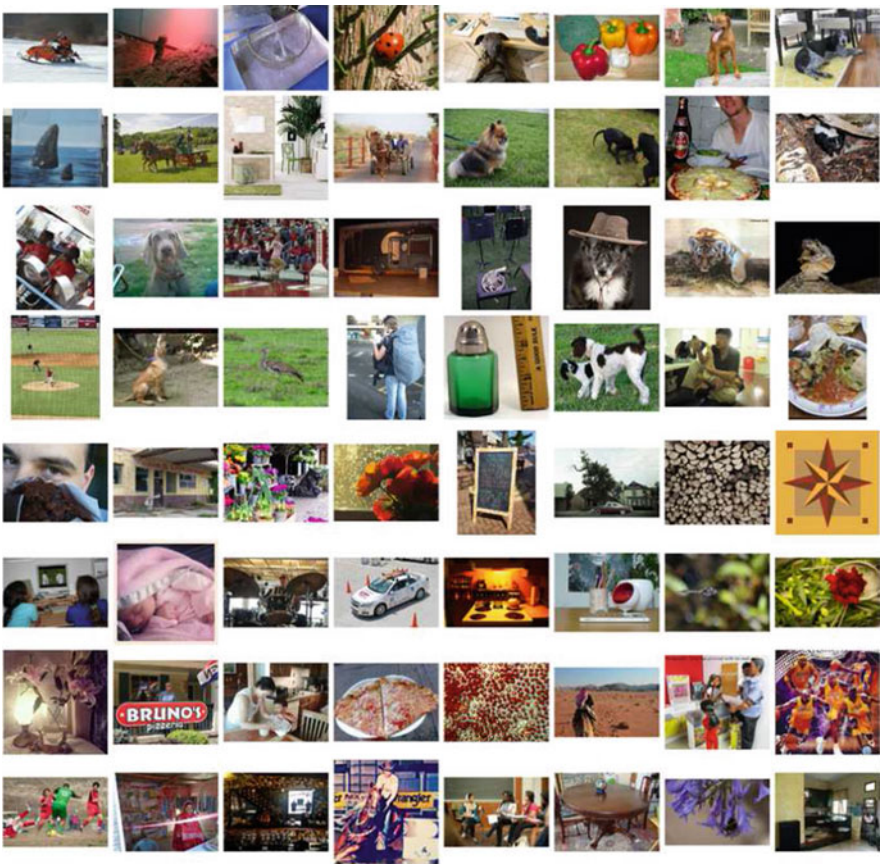

Fig. 2.7 Examples in the ImageNet dataset

2.2.5 Scene Understanding (SUN) Dataset

The Scene understanding (SUN) dataset, introduced by Xiao et al., finds applications in many research fields, such as scene recognition, computer vision, human perception, cognition and neuroscience, machine learning, data mining, computer graphics and robotics research. The SUN dataset contains 899 categories and 130,519 images [14] in total. It has been widely used in various scene related computer vision researches, such as scene classification [14], scene recognition [15], and indoor/outdoor image classification [16]. It was motivated by the demand to build a rich and diverse dataset that includes our daily experienced scenes in the real world as much as possible. Being different from the object detection datasets such as the PASCAL [11] and the Caltech 256 category datasets [17], images in the SUN dataset are all about scenes where human can navigate or interact with. The SUN dataset

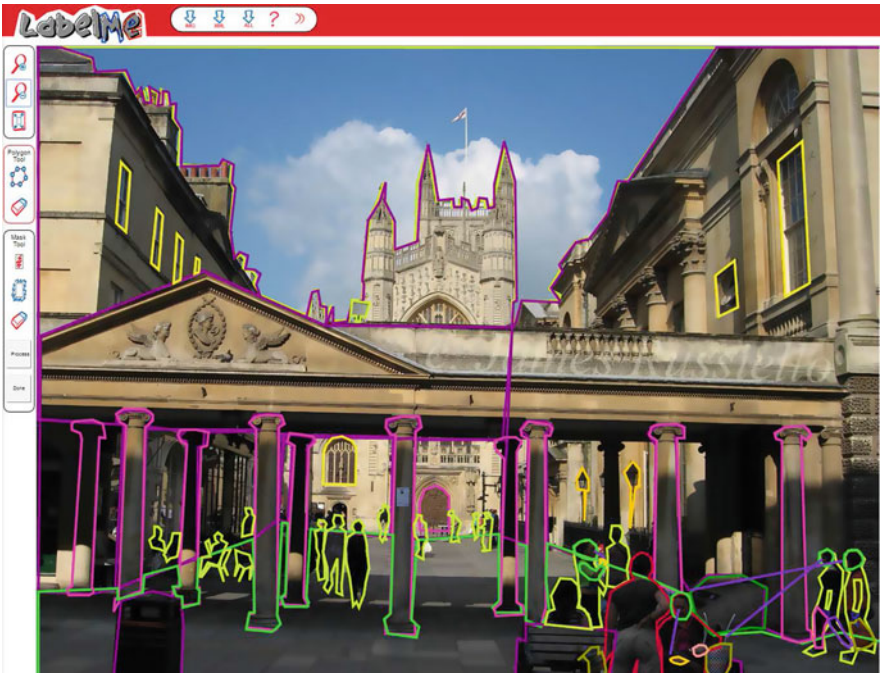

Fig. 2.8 An example of annotations in LabelMe

is currently the largest scene dataset in terms of the image number and the scene number.

The scene category of the SUN dataset is huge. We can easily think of some scene categories such as the coast, the field, the meeting room etc. Is Grand Canyon a scene? Should it be a category? How can we include as many scene terms as possible? The category terms are selected from the 70,000 terms of the WordNet [9] used in the tiny images dataset [8]. These terms describe scenes, places and environments. There are several criteria in selecting scene category terms. First, places terms that are too broad to evoke a specific visual identity (such as territory, workplace and outdoors) and places names (such as Grand Canyon or New York) are not included. Second, specific types of objects which are scene related are included, such as buildings (skyscraper, house and hangar) which makes the scene categories more diverse. Third, it contains specific domains such as the pine forest, rainforest and orchard which all belong to the wooded area. Besides the terms in the WordNet, a few categories missed by the WordNet are added. After the first round of term selection, there are about 2500 initial terms. After merging terms by synonyms and separating scenes of different visual identities (indoor and outdoor views of churches), there are 899 categories. This is far more than the previously created 8-scene dataset [1] and the 15-scene dataset [4]. We can see that the SUN dataset really contains comprehensive scene categories in Fig. 2.9.

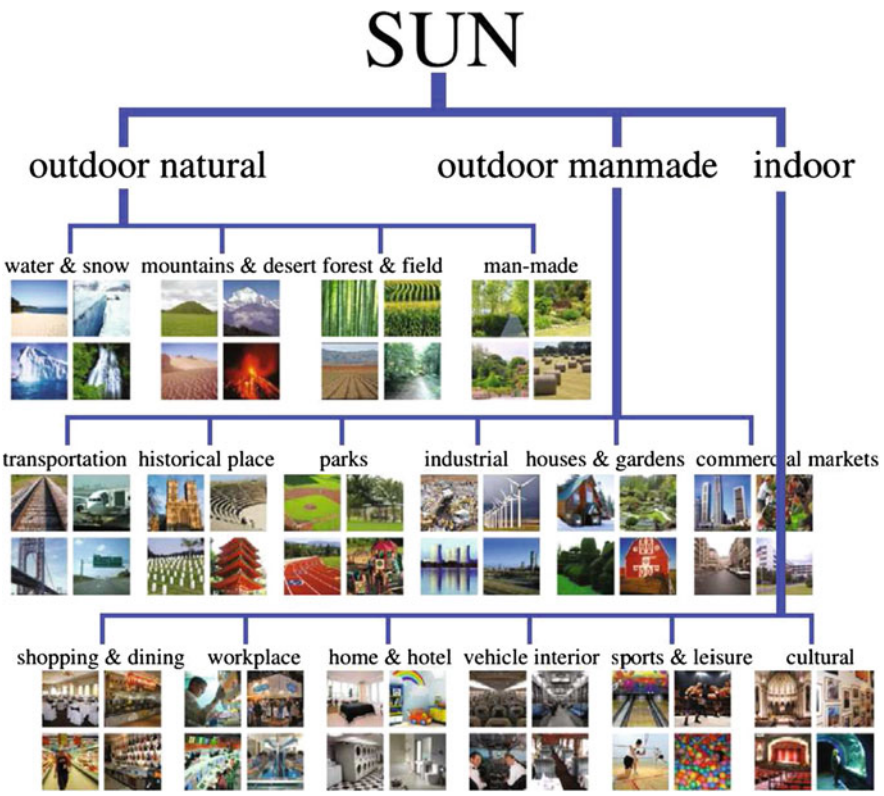

Fig. 2.9 Visualization of scene hierarchies in SUN

There is a subset of the SUN 899 categories dataset containing 397 categories. It is more popular in the research field since each category contains at least 100 images. Other categories not included in the 397 category contain fewer images. The total number in the SUN database is large. Images in the SUN dataset come from the Internet search. In each category of the 8-scene and the 15-scene dataset contains hundreds of images, which is far less than that in the SUN dataset. In contrast with the eight scene and the fifteen scene datasets, images in each category of the SUN dataset are more diversified. As a result, this dataset imposes more challenges on the scene classification and recognition tasks.

As compared with the 80 million tiny image dataset, images in the SUN dataset are of much higher resolution. Images in the SUN dataset have a resolution of at least 200×200 pixels. Degenerate or unusual images (black and white, distorted colors, very blurry or noisy, incorrectly rotated, aerial views, noticeable borders) were removed in the image collection process.

[14] conducted experiments to compare human scene classification accuracy and machine classification accuracy. The purpose was to show that the SUN dataset was

constructed consistently with minimal overlap between categories and that the scene classification task would be a difficult one.

To facilitate participants to understand the 397 scene categories, category terms are grouped in a three-level tree. The category on the leaf-level is the most specific one. The participants can navigate through the three-level hierarchy tree to reach a specific scene type. Each leaf-level SUN category interface shows an example of the image to help participants better know what the image looks like in each category.

Human scene classification accuracy was measured on 20 distinct test scenes in each category by Amazon's Mechanical Turk (AMT). There are $397 \times 20 = 7940$ experiments or HITs (Human Intelligence Tasks in AMT parlance). Human take 61 s per experiment and can achieve 58.6 % accuracy at the leaf level on the average. Considering the large number of categories, human classification accuracy is already very high. There are "good works" that can achieve accuracy as high as 95 % on the relatively easy first level of the hierarchy and the leaf-level accuracy rises to 68.5 %. One author involved in constructing the SUN dataset achieved 97.5 % at the first-level and 70.6 % at the leaf-level. The "good works" are trustworthy. By analyzing the confusion scene categories, these confused scenes are semantically similar and they are restricted to only a few scenes. The experiment shows that human classification accuracy is far below 100 % at the leaf-level. Image classification is actually a difficult task for human, and it is even more difficult for the computer.

Twelve individual methods are compared in scene classification, including GIST [1], HOG2x2 [18], Dense SIFT [4], LBP [19], Sparse SIFT histograms [20, 21], SSIM [20], Tiny Images [8], Line Features [22, 23], Texton Histograms [24], Color Histograms, Geometric Probability Map [25] and Geometric specific histograms [26]. The extracted features are all relevant to scene classification. The classifier is trained using the one-vs-remaining SVM. To compare the performance on different datasets, experiments are also conducted on the 15-scene dataset [1, 3, 4]. The results are shown in Fig. 2.10. The performance curve labeled by "all" is to adopt the weighted sum of individual features as the new feature. The weight is chosen as the proportion to the fourth power of its individual accuracy.

From the results, we see that the correct classification rate of each individual feature ranges from 50–82 % while that of the "all" curve can perform up to 88.1 % in the scene recognition task for the fifteen scene dataset. However, when being applied to a much larger dataset, i.e. the SUN dataset, the performance of each individual feature drops to 5–28 % while the "all" curve drops to 38 %. This shows that, when the number of categories increases, the problem becomes more difficult.

In the work [14], we also see the differences between human and computer classification errors. Human errors mostly lie in semantically similar categories. However, computer errors are due to wrong features. The computer can make errors on semantically unrelated scenes. On the other hand, the computational methods are more accurate than human [14] for some categories.

A scene classification task is to classify scenes into different categories. A more difficult problem is scene detection. Being analogous with the object detection problem that identifies an sub-image as a certain object class, the scene detection problem is to identify a scene inside in an image. This problem arises from the observation

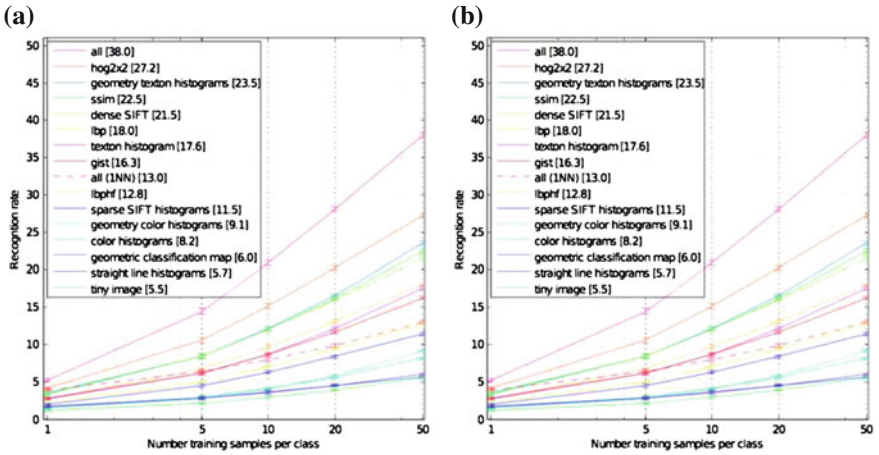

Fig. 2.10 Scene recognition with 24 scene types. **a** 15-scene dataset. **b** SUN dataset

that the real world scene might not be well divided into scenes of fifteen categories where each image exactly belongs to one scene. There might be several scene types in one image. To conduct the scene detection task, 24 well-sampled categories of the 398 categories are used in training. Another 104 photographs containing an average of 4 scene categories are used as test images. An example is shown in Fig. 2.11. Although only 24 scene categories are trained as possible scenes in the test image, the scene detection accuracy is not high, range from 8–66 % even using all features to train the classifier [14].

Fig. 2.11 **a** A photograph that contains multiple scene-types, **b** the ground truth annotations, **c–e** Detections of the beach, the harbor, and the village scene categories in one single image. In all visualizations, correct detections are in *green* and incorrect detections are in *red*. The bounding box size is proportional to classifier confidence. For this and other visualizations, all detections above a constant confidence threshold are shown. In this result, the harbor detection is incorrect since it does not overlap much with the ground truth "harbor" annotation while "beach" and "village" are acceptable

Both the scene classification performance comparison between human and computer and the performance of scene classification and detection using SVM show the difficulty of the scene understanding problem. A large scale dataset adds challenges to the problem. However, this database is needed to evaluate various algorithms and stimulate more advanced algorithms to be developed in the near future.

2.2.6 Places205 Dataset

With more and more data available and the emergence of deep learning trends, researchers stared to prepared even more larger dataset than SUN. Place205 [27] is the latest and most challenging one. Places205 contains 2,448,873 images from 205 scene categories in total. It is treated as the largest scene classification dataset, and mainly prepared for the purposes of Convolution Neural Network (CNN) training.

Comparing with ImageNet and SUN, in Fig. 2.12, Place205 shows extreme data abundance, which is very crucial for discriminative model learning for CNN with deep structures that has millions of parameters. In paper [27], author trained the huge CNN network with 2,448,873 image in 6 days and present superior results on traditional datasets with the trained deep features. In Fig. 2.13, we see the example output of Places-CNN for the query image on the left.

Since Places205 is proposed quite recently, and only CNN can take advantages of the such a large number of training images, there is little work benchmark using this dataset. However, it brought out a new definition of "large-scale image understanding".

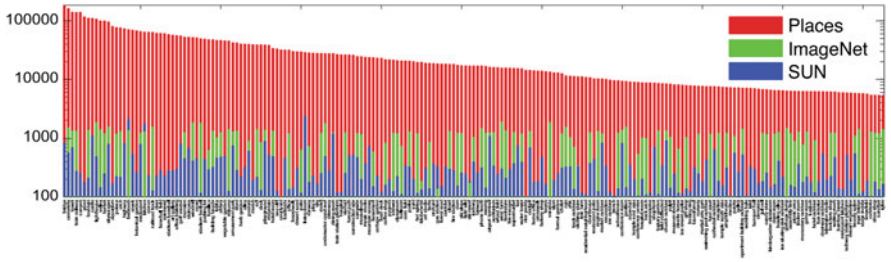

Fig. 2.12 Comparison of the numbers of images in Places 205 with ImageNet and SUN. Note that ImageNet only has 128 of the 205 categories, while SUN contains all of them. To compare them, we select a subset of Places. It contains the 88 common categories with ImageNet such that there are at least 1000 images in ImageNet. We call the corresponding subsets SUN 88 and ImageNet 88

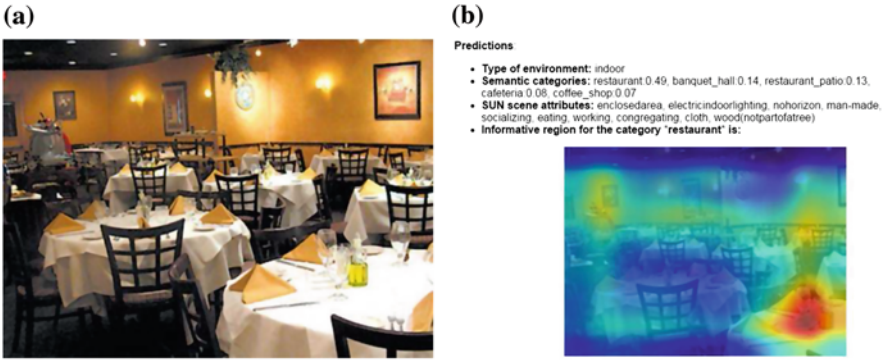

Fig. 2.13 Demo of the trained Places-CNN model: given a query image on the *left*, classification results are given with multiple classification soft scores

References

1. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vision **42**(3), 145–175 (2001)
2. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: label transfer via dense scene alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 1972–1979. IEEE (2009)
3. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 2, pp. 524–531. IEEE (2005)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 2169–2178. IEEE (2006)
5. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, pp. 1–8. IEEE (2007)
6. Hoiem, D., Efros, A., Hebert, M., et al.: Geometric context from a single image. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005, vol. 1, pp. 654–661. IEEE (2005)
7. Hoiem, D., Efros, A., Hebert, M., et al.: Closing the loop in scene interpretation. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
8. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
9. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
10. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step (2004)
11. Everingham, M., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 248–255. IEEE (2009)
13. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vision **77**(1–3), 157–173 (2008)

14. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
15. Gao, T., Koller, D.: Discriminative learning of relaxed hierarchy for large-scale visual recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2072–2079. IEEE (2011)
16. Pavlopoulou, C., Yu, S.X.: Indoor-outdoor classification with human accuracies: image or edge gist? In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 41–47. IEEE (2010)
17. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
19. Ojala, T., Pietiknen, M., Mnp, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
20. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vision Comput. **22**(10), 761–767 (2004)
21. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 1, pp. I-488–I-495, IEEE (2004)
22. Hays, J., Efros, A.: Im2gps: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
23. Zhang, J.K.W.: Video compass. Computer Vision? ECCV 2002, pp. 476–490. Springer (2002)
24. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings, vol. 2, pp. 416–423. IEEE (2001)
25. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. Int. J. Comput. Vision **75**(1), 151–172 (2007)
26. Lalonde, J.F., Hoiem, D., Efros, A.A., Rother, C., Winn, J., Criminisi, A.: Photo clip art. ACM Trans. Graph. **26**(3), 3 (2007)
27. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)

http://www.springer.com/978-981-10-0629-6

Big Visual Data Analysis
Scene Classification and Geometric Labeling
Chen, C.; Ren, Y.; Kuo, C.-C.J.
2016, X, 122 p. 94 illus., 12 illus. in color., Softcover
ISBN: 978-981-10-0629-6