Scene Recognition via Bi-enhanced Knowledge Space Learning

Jin Zhang[1], Bing-Kun Bao[2,3(✉)], and Changsheng Xu[1,3]

[1] Hefei University of Technology, Hefei, China
jin_zhang@duckj.cn
[2] Nanjing University of Posts and Telecommunications, Nanjing, China
bingkunbao@njupt.edu.cn
[3] National Lab of Pattern Recognition, Institute of Automation, Beijing, China
csxu@nlpr.ia.ac.cn

Abstract. Scene recognition is one of the hallmark tasks in computer vision, as it provides rich information beyond object recognition and action recognition. It is easy to accept that scene images from the same class always include the same essential objects and relations, for example, scene images of "wedding" usually have bridegroom and bride next to him. Following this observation, we introduce a novel idea to boost the accuracy of scene recognition by mining essential scene sub-graph and learning a bi-enhanced knowledge space. The essential scene sub-graph describes the essential objects and their relations for each scene class. The learned knowledge space is bi-enhanced by global representation on the entire image and local representation on the corresponding essential scene sub-graph. Experimental results on the constructed dataset called Scene 30 demonstrate the effectiveness of our proposed method.

Keywords: Scene recognition · Sub-graph mining · Bi-enhanced

1 Introduction

Scene recognition is one of the most challenging tasks in image classification and various scene recognition methods have been proposed over the past decades [3,12,15,21–23,25,26,29]. To deal with large intra-class variance caused by nuisance factors such as pose, viewpoint and occlusion, it normally requires two stages for a scene recognition solution, that is, scene representation and scene classification.

Scene representation aims to fully use all the information of scene images to extract discriminative features. It explores not only the generalized characteristics in the same category but also the distinctive characteristics among different categories. The representation methods can be mainly classified into two categories, hand-crafted and deeply-learned representations. In early studies, hand-crafted representation was popular due to its simplicity and low computational cost. These methods only capture low-level information, such as texture and

C.-Y. Chang et al. (Eds.): ICS 2018, CCIS 1013, pp. 215–223, 2019.
https://doi.org/10.1007/978-981-13-9190-3_23

structure of the information. In recent works, deeply-learned feature extraction methods exploit high-level semantic information in scene images by using Convolutional Neural Networks (CNNs).

In this paper, we propose an effective scene recognition framework, which firstly extracts the essential scene sub-graph for each scene class, then learns a classifier to distinguish different scene classes by learning a bi-enhanced knowledge space. The whole work is based on the scene images and their corresponding scene graphs. The main contributions of our work are summarized as follows:

- We propose a novel framework to extract discriminative representation from both entire image and essential scene sub-graph for scene recognition. The learned bi-enhanced knowledge space is proved to be useful for classification.
- This work explores a pioneer study on learning knowledge graph, i.e. essential scene sub-graph, for scene recognition. The proposed approach has great potential for other categorization tasks, while enables people to think about how knowledge graph can better drive current tasks.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. The proposed framework including the essential scene sub-graph mining and the bi-enhanced knowledge space learning is described in Sect. 3. Experimental results are reported and discussed in Sect. 4, followed by the conclusion in Sect. 5.

2 Related Work

In this section, we briefly review the related work on scene representation and scene classification.

Scene representation is the most important step in scene recognition task, which aims to extract discriminative features from scene images. GIST [15], which is one of hand-crafted global features, lexicographically converts an entire scene image into a high-dimensional feature vector, but fails to exploit local structure information in scenes, especially the indoor scenes with complex spatial layouts. Methods focusing on local features, such as OTC [14] and CENTRIST [22], firstly describe the structure pattern of each local patch and then combine the statistics of all patches into a concatenated feature vector. Recently, as Convolutional Neural Networks (CNNs) have made remarkable progress on image recognition, deeply-learned methods have been widely adopted. Gong *et al.* [7] proposed a multi-scale orderless pooling (MOP) method to extract fully-connected features on image local patches. While these methods have achieved encouraging performance, a largely overlooked aspect is the role of the scale and its relation with the feature extractor in a multi-scale scenario. Herranz *et al.* [8] adapted the feature extractor to each particular scale, which combined ImageNet-CNNs [17] and Places-CNNs [29] to improve classification performance. However, the essential objects and their relations are still not fully utilized, while much information extracted from patches is redundant. Furthermore,

most of the recent methods need to produce the proposal of each objects, which push the computational costs too high when dealing with large scale dataset.

Over the past decades, many methods have been proposed for scene classification [2,6,16,19,20,27,28] and can be categorized into two groups: generative models and discriminative models. Generative models usually adopt hierarchical Bayesian to express various relations in a complex scene, such as Markov random fields (MRF) [6], hidden Markov model (HMM) [19] and latent Dirichlet allocation (LDA) [1]. However, these models need to build complex probabilistic graph model and require high computational cost. The discriminative models extract feature descriptors from images and then encode them into a fixed length vector for classification. The typical classifiers include logistic regression and support vector machine (SVM) [2]. Especially, the SVM classifier has been widely used for scene classification. Object bank (OB) [13] and deformable part based model (DPM) [5] are representative examples of training a feature classifier for scene classification. Unlike the generative models, the parameters of discriminative models are easy to learn for feature classification.

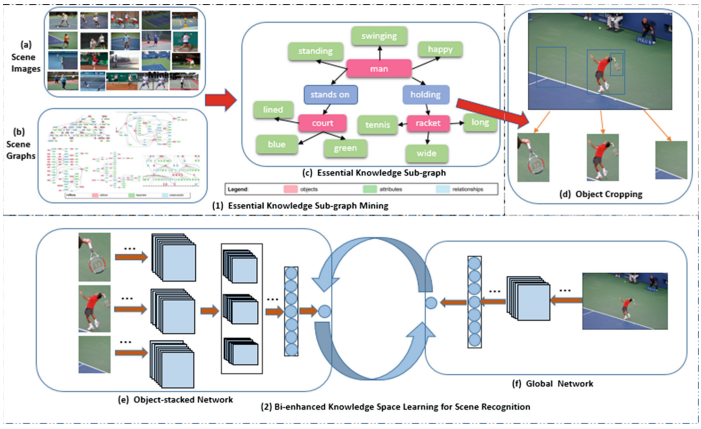

Fig. 1. Overview of proposed framework. The model consists of: (1) essential scene sub-graph mining; (2) bi-enhanced knowledge space learning for scene recognition.

3 Our Approach

Our proposed framework is illustrated in Fig. 1, which contains two key stages: essential scene sub-graph mining and bi-enhanced knowledge space learning. Firstly, we adopt a statistical method to mine the essential scene sub-graph for each scene class. Next, the bi-enhanced knowledge space is sought for scene image recognition by iteratively learning representations from essential scene sub-graph and entire image. In this section, we present the details of the proposed framework.

3.1 Essential Scene Sub-graph Mining

The scene graph is a graph of each scene image to describe all the objects, attributes and inter-object relations. Our approach attempts to mine the essential scene sub-graph by using the similarity between the scene graphs from the same class.

For essential scene sub-graph mining, we statistically analyze the frequencies of objects for each scene. Firstly, we count the occurring frequencies of all object sets for each scene class. Next, we choose object sets with the highest frequencies and size varying from 1 to 6 for each scene class. Lastly, we calculate the percentages of images including all the objects in above selected object sets for each scene class, and then the average of them for all the scene classes. Taking the scene of "tennis game" as an example, after counting the occurring frequencies of all object sets in all "tennis game" scene images, we obtain that *tennis player* surfacing out when object set size is 1 and 98.5% of images include it. Similarly, *tennis player, tennis court* is selected when object set size is 2, with 76.4% of images include them. More details on essential scene sub-graph mining are shown in Algorithm 1.

Algorithm 1. Essential Scene Sub-graph Mining

Input: Image set C_j in the j-th scene class.
Output: Essential scene sub-graph(objects set $\hat{O}_j$ that contains relations) for the j-th scene class

Initiate $k[m][m] = 0$ and a empty dictionary D_j
for $i = 0$ to $N(C_j)$ **do**
  while c_i has $object$ **do**
    $S_i.add(object)$
  end while
  for object in S_i **do**
    if $object$ not in D_j.Keys() **then**
      $D_j[object] = D_j[object] + 1$
    else
      $D_j[object] = 1$
    end if
  end for
end for
pick $O_m = \{o_1, o_2, , , o_m\}$ from the top of D_j
for $n = 0$ to $N(C_j)$, $i = 0$ to m, $l = 0$ to m **do**
  if (o_i,o_j) has edge **then**
    $k[i][l] = k[i][l] + 1$
  end if
end for
if $set(i, l, p, q) == 3$ **then**
  return $\hat{O}_j$=(o_1,o_2,o_3),(o_2 is the repeat object)
else
  return $\hat{O}_j$=(o_i,o_j,o_p)
end if

3.2 Bi-enhanced Knowledge Space Learning

This section aims to illustrate the learning of a knowledge space that saves useful and discriminative features from entire images and essential scene sub-graph. The structure of the whole model is shown in Fig. 2. It includes three parts: (1) object-stacked network, which learns features from essential scene

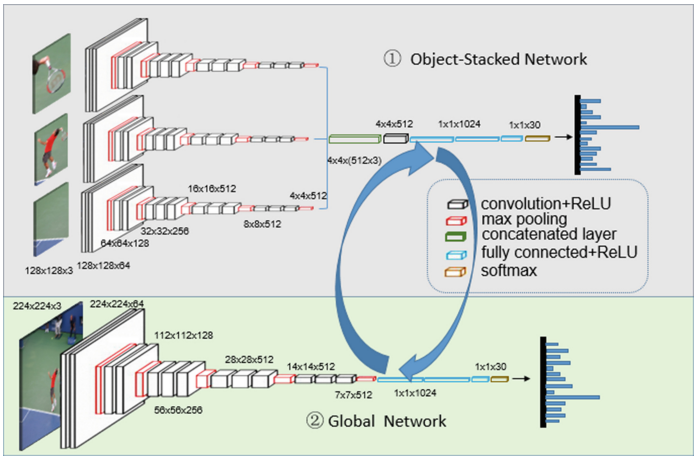

Fig. 2. Illustration of the bi-enhanced knowledge space learning. ① is the object-stacked network and ② describes the global network. The whole figure demonstrates an iterative process for knowledge space learning.

sub-graph enhanced by global representation, (2) global scene network, which learns features from the entire image enhanced by object-stacked representation, and (3) bi-enhanced knowledge space optimization, which iteratively seeks the knowledge space from both object-stacked representation learning and global representation learning.

Inspired by Huang *et al.* [9] and considering the structure of essential scene sub-graph, we adopt an object-stacked network to process three objects and the relations in essential scene sub-graph as shown in Fig. 2. The object-stacked network contains three separate convolutional blocks, a concentrated layer which is adopted to combine the three-stream features, a 1×1 convolutional layer which is to reduce dimension, and a fully-connected layer which is utilized to build a knowledge space. The objective function is in Eq. (1):

$$\min_{W,b} \sum_{i=1}^{m} (||f(o_{i_1}, o_{i_2}, o_{i_3}) - h(c_i)||) + \lambda ||W||^2 \tag{1}$$

where W and b are the weight and bias of the layers in network, respectively, m is the number of all the scene images, $f(\cdot)$ is the output of the first fully-connected layer $f6$ from object-stacked network, $h(c_i)$ is the global representation of image c_i which is learned from global network. o_{i_1}, o_{i_2}, o_{i_3} are the objects of essential scene sub-graph cropped from image c_i and $h(\cdot)$ is the output of the first fully-connected layer from global scene network, $\lambda ||W||^2$ is regularization term. Note that the object o_{i_2} which has relations to other two objects o_{i_1}, o_{i_3} is fed into the second stream. For example, for the scene of "tennis game", the essential objects are *man*, *court* and *racket*. The relations from essential scene sub-graph are *the man holding the racket* and *the man stands on the court*. Obviously, *man* has

the relations to both *court* and *racket*, and is inputted into the second stream. If the image does not contain all three essential objects, we set the value of missing object as 0.

Recall that our task is a classification problem, we add another 2 fully-connected layers and softmax layer after fc6. The final objective function is expressed in Eq. (2)

$$\min_{W,b} \gamma \sum_{i=1}^{m} ||f(o_{i_1}, o_{i_2}, o_{i_3}) - h(c_i)|| + \xi\lambda||W||^2 - \delta \sum_{i=1}^{m} (y_i \log(T(o_{i_1}, o_{i_2}, o_{i_3})) \quad (2)$$

where ξ is used to determine whether to join regularization, γ and δ are the parameters introduced to reduce the difference between the two losses that we set 0.01 and 1. y_i is the label of image c_i, $T(\cdot)$ is the output of final softmax layer in object-stacked network.

Similar to object-stacked network, we adopt a CNN model to learn global representation as shown in Fig. 2. It contains five convolutional blocks, two fully-connected layers and a softmax layer for classification. The dimension of the first fully-connected layer is equal to the dimension of representation from object-stacked network. The objective function is shown in Eq. (3):

$$\min_{W,b} \alpha \sum_{i=1}^{m} ||h(c_i) - f(o_{i_1}, o_{i_2}, o_{i_3})|| + \mu\lambda||W||^2 - \beta \sum_{i=1}^{m} (y_i \log(H(c_i)) \quad (3)$$

where $H(\cdot)$ is the output of final softmax layer in global scene network. The parameters α and β are utilized to balance these two losses, and μ controls whether to use regularization term. The meaning of the remaining parameters is the same as before. We use mini-batch stochastic gradient descent (SGD) to optimize Eq. (3). When Eq. (3) reaches an optima, we obtain the global representation enhanced by object-stacked representation.

Based on the above mentioned two networks, an iterative process between them is adopted. The iterative process is initiated by object-stacked network with cross-entropy cost function instead of global representations. Next, at each iteration, we update object-stacked representations by optimizing Eq. (2) which is enhanced with global representations, and then adjust global representations by optimizing Eq. (3) which is enhanced with object-stacked representations. The knowledge space is optimized iteratively until convergence. For test, we only employ the trained global network to predict the scene class.

4 Experiments

This section demonstrates the effectiveness of the learned bi-enhanced knowledge space on Scene 30.

4.1 Datasets and Implementation

To better demonstrate the proposed method in large scale dataset, we construct Scene 30 from Visual Genome [10]. The constructed Scene 30 contains 4608 color

images of 30 different scenes including both indoor and outdoor scenes. The number of images varies across categories with at least 50 images per category. Each image has a corresponding scene graph. There are 10,034 objects and 30,000 types of relations in total in Scene 30. We split 85% of each class from the entire dataset for training and the rest as test set. The object-stacked network and global scene network are implemented using the open-source package Keras [4]. We adopt the VGG-16 model pre-trained in ImageNet [18]. In object-stacked network, the cropped object patches are resized to 128×128, and the input of global scene network are warped to a 224×224. The features of scene and object-stacked network are extracted from the layer of $fc6$.

4.2 Result and Comparison

Table 1 shows the comparison results. From the table, we can see that the accuracy of the classification increased from 82.51% to 88.29% after two iteration cycles. Moreover, through the bi-enhanced knowledge space learning, global network and object-stacked network capture more meaningful and discriminative information. The accuracy of the classification in object-stacked network increased from 89.60% to 90.32%. Similarly, the accuracy of the global network also increased from 82.51% to 86.71% and then to 88.29% under the supervision of local essential objects features.

Table 1. Recognition performance comparisons in different iterations

Methods	Accuracy
VGG-16 [18]	82.51%
OSN-iter1	89.60%
OSN-iter2	90.32%
OSN-iter2 fc6 + SVM	87.57%
GN-iter1	86.71%
GN-iter2 + SVM	87.43%
GN-iter2	88.29%

Table 2. Recognition performance comparisons on Scene 30

Methods	Accuracy
AlexNet [11]	72.40%
VGG-16 [18]	82.51%
PlaceNet205 (AlexNet) [29]	72.19%
PlaceNet365 (AlexNet) [24]	76.56%
PlaceNet205 (VGG-16) [29]	86.77%
PlaceNet365 (VGG-16) [24]	86.67%
HybridNet (AlexNet) [29]	73.54%
HybridNet (VGG-16) [24]	87.08%
Our fc6 + SVM	87.43%
Ours	**88.29%**

We then evaluate our proposed methods on Scene 30 and compare it with several recent CNN based methods. Table 2 records the recognition accuracy of our approach and other methods where we achieve the highest recognition rate. The method named "Our $fc6$ + SVM" extracts the feature in $fc6$ and trains a SVM for classification. The method named "Ours" directly utilizes the global network to predict the scene class. From the table, we have following 3 observations. (1) VGG-16 outperforms AlexNet. For example, VGG-16 in PlaceNet 365

is 10.11% higher than AlexNet. Therefore, we choose VGG-16 as a basic model. (2) the essential scene sub-graph is beneficial to scene recognition. The accuracy of our approach is 1.52% higher than PlaceNet 365 (VGG-16) and 1.21% higher than HybridNet 1365 (VGG-16). (3) The logistic regression is better than SVM for scene classification. We analyze that our model is an end to end framework for testing, while the SVM extracts the $fc6$ feature and then is optimized for classification.

5 Conclusion

In this paper, we propose a novel framework to learn the discriminative representations from both entire scene image and essential scene sub-graph. In future work, we will focus on utilizing the probability graph model to mine the essential scene sub-graph, such as Markov random fields (MRF [6]), and build a more accurate relationship between the scene image and objects.

Acknowledgement. This work is supported by the National Key Research & Development Plan of China (No. 2017YFB1002800), by the National Natural Science Foundation of China under Grant 61872424, 61572503, 61720106006, 61432019, and by NUPTSF (No. NY218001), also supported by the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, and the K.C. Wong Education Foundation.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Chen, P.H., Lin, C.J., Schölkopf, B.: A tutorial on-support vector machines. Appl. Stoch. Models Bus. Ind. **21**(2), 111–136 (2005)
3. Cheng, X., Lu, J., Feng, J., Yuan, B., Zhou, J.: Scene recognition with objectness. Pattern Recogn. **74**, 474–487 (2018)
4. Chollet, F., et al.: Keras (2015). https://github.com/keras-team/keras
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection-with discriminatively trained part-based models. PAMI **32**(9), 1627–1645 (2010)
6. Geman, S., Graffigne, C.: Markov random field image models and their applications to computer vision. In: Proceedings of the International Congress of Mathematicians, vol. 1, p. 2 (1986)
7. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 392–407. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_26
8. Herranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: CVPR, pp. 571–579 (2016)
9. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked CNN for fine-grained visual categorization. In: CVPR, pp. 1173–1182 (2016)
10. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017)

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Bao, B.-K., Zhu, G., Shen, J., Yan, S.: Robust image analysis with sparse representation on quantized visual features. IEEE Trans. Image Process. **22**(3), 860–871 (2013)
13. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: Advances in Neural Information Processing Systems, pp. 1378–1386 (2010)
14. Margolin, R., Zelnik-Manor, L., Tal, A.: OTC: a novel local descriptor for scene classification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 377–391. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_25
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representationof the spatial envelope. IJCV **42**(3), 145–175 (2001)
16. Parizi, S.N., Oberlin, J.G., Felzenszwalb, P.F.: Reconfigurable models for scene recognition. In: CVPR 2012, pp. 2775–2782. IEEE (2012)
17. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Stamp, M., Professor, A.: A revealing introduction to hidden Markov models. IEEE ASSP Magruine **1**(24), 258–261 (2004)
20. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: ICCV 2005, vol. 2, pp. 1331–1338. IEEE (2005)
21. Wang, Z., Wang, L., Wang, Y., Zhang, B., Qiao, Y.: Weakly supervised patchnets: describing and aggregating local patches for scene recognition. TIP **26**(4), 2028–2041 (2017)
22. Wu, J., Rehg, J.M.: Centrist: a visual descriptor for scene categorization. PAMI **33**(8), 1489–1501 (2011)
23. Xie, G.S., Zhang, X.Y., Yan, S., Liu, C.L.: Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. IEEE Trans. Circuits Syst. Video Technol. **27**(6), 1263–1274 (2017)
24. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. PAMI **40**, 1452–1464 (2017)
25. Bao, B.-K., Liu, G., Changsheng, X., Yan, S.: Inductive robust principal component analysis. IEEE Trans. Image Process. **21**(8), 3794–3800 (2012)
26. Bao, B.-K., Min, W., Li, T., Changsheng, X.: Joint local and global consistency on interdocument and interword relationships for co-clustering. IEEE Trans. Cybern. **45**(1), 15–28 (2015)
27. Min, W., Bao, B.-K., Mei, S., Zhu, Y., Rui, Y., Jiang, S.: You are what you eat: exploring rich recipe information for cross-region food analysis. IEEE Trans. Multimed. **20**(4), 950–964 (2018)
28. Bao, B.-K., Changsheng, X., Min, W., Hossain, M.S.: Cross-platform emerging topic detection and elaboration from multimedia streams. TOMCCAP **11**(4), 54 (2015)
29. 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)