DEEP-CSSR: SCENE CLASSIFICATION USING CATEGORY-SPECIFIC SALIENT REGION WITH DEEP FEATURES

Mengshi Qi, Yunhong Wang

Beijing Key Laboratory of Digital Media
School of Computer Science and Engineering
Beihang University
Beijing, China, 100191

ABSTRACT

Researches in neuroscience and biological vision have shown that the bio-inspired methods have excellent recognition performance, such as the salient detection, artificial neural network and the ganglion cell inspired image feature. In this paper, we introduce a novel framework towards scene classification using category-specific salient region(CSSR) with deep CNN features, called Deep-CSSR. Firstly, by using the salient region detection algorithm, we extract a set of image patches which contain the salient regions. Also we apply DERF, a novel bio-inspired image descriptor, to represent the salient patches and clustering all of them to remove the outliers. Then we learn the CSSR filters and construct the CSSR representation. Further more, we do scene image classification using CSSR representation concatenate with the deep CNN features extracted from the whole images. By using this new pipeline, we obtain better results than recent methods over MIT Indoor 67 and Sun397 databases.

Index Terms— Scene Classification, Salient Region Detection, Deep CNN features

1. INTRODUCTION

Owing to much variation of objects and layouts from different views, and the difference of spatial information between scene categories, the performance of scene classification can not be satisfied completely in academia and industry. Fortunately, we can seek out a better solution from visual and attention research which inspired by biology.

From the previous study of scene classification and recognition, scene categories can be modeled directly from low-feature, such as SIFT[1] and DAISY[2]. However, these low-feature descriptors can not take consideration of the information processing mechanism of the human visual system. So these low features are difficulty to obtain semantic information and whole factors of scene categories. It is necessary to find a distinctive efficient robust feature, which in accordance with recognition process of the visual cortex ganglion cells.

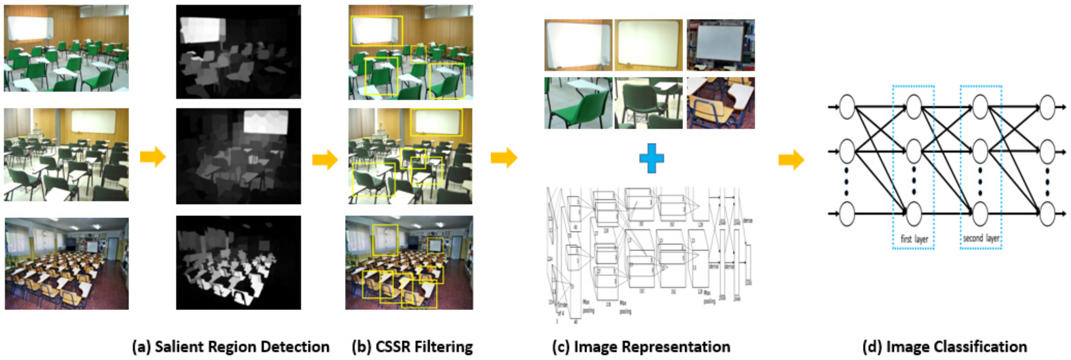
Some recent approaches adopt mid-level feature that are extracted from scene image patches. And there are two general ways to achieve the image patches: randomly sampled patches and densely sampled patches[3]. But these two methods have obvious disadvantages: the randomly sampled pathes can not reflect the full information of images and dense sampled patches always make much computation and storage consumption. Also these methods choose semantic patches or regions to build mid-level vocabularies must be labeled by human labor. It is well known that human have strong ability to recognize the distinguish regions which have high saliency in images quickly and correctly. For instance, television and sofa are salient stuffs in living room scenes, chairs and blackboard are salient stuffs in classroom scenes. So unsupervised salient detection would be a preeminent approach.

Also, the deep convolution neural network[4] makes a breakthrough in artificial intelligence and pattern recognition. The deep CNN gets the deep features which have a great capability for representing the images and objects through training. But the present methods still extract holistic CNN features of full images and train common Alex's net[4]. They not only regardless the layout and structure information of scene images, but also can not extract the efficient and specific features of the scene images. Therefore it is essential to construct a category-specific representation of scene image with deep CNN features.

In our paper, we propose a novel framework using category-specific salient region with deep features(DEEP-CSSR) to do scene classification. Firstly, we adopt a context-aware salient detection algorithm to extract salient region patches in unsupervised way. Then we do clustering for these salient regions using DERF[5] feature descriptor, which is inspired by biological modeling of the P ganglion cells. After that, we learn category-specific salient region(CSSR) filters with appearance and position of CSSR. At last, we construct the CSSR representation of images, and concatenate with the deep CNN features of the whole images to do classification by neural network.

The rest of the paper is organized as follows. Our ap-

978-1-4673-9961-6/16/$31.00 ©2016 IEEE



(a) Salient Region Detection (b) CSSR Filtering (c) Image Representation (d) Image Classification

Fig. 1. Overview of our framework.

proach for scene classification is proposed in section 2. Then the experimental results are shown in Section 3. At last the conclusion and future work in Section 4.

2. APPROACH

Our approach has four major steps: salient region detection, category-specific salient region filtering, followed by scene image representation and classification. Our framework is shown in Figure 1.

2.1. Salient Region Detection

As for the scene classification, it is necessary to take the context of the dominant regions and objects into consideration. So we adopt an unsupervised context-aware saliency detection method from [6].

Saliency Detection At first, we define an image patch centered at pixel i which its appearance different from others as the salient region. So our algorithm computes saliency by measuring the dissimilarity of each image patch from the same image. We adopt two kinds of the Euclidean distance, that is, $d_{position}(x^i, x^j)$ is the Euclidean distance between the position of patch x^i and x^j, and $d_{color}(x^i, x^j)$ which represents the Euclidean distance between the patch x^i and x^j in CIE $L*a*b$ color space. Depend on these position relationship, we can measure the dissimilarity of two patches as

$$d(x^i, x^j) = \frac{d_{color}(x^i, x^j)}{1 + c * d_{position}(x^i, x^j)} \quad (1)$$

where the constant $c = 3$, and $d_{position}(x^i, x^j)$, $d_{color}(x^i, x^j)$ are normalized to the range $[0, 1]$ in this formulation.

Generally, we find N most similar image patches, and if one image patch x^i is highly different from the other similar image patches, it means that x^i has so high saliency that human attention would be concentrated on. So we can define the saliency value S^i of x^i in single image as

$$S^i = 1 - exp\left\{-\frac{1}{N}\sum_{n=1}^{N} d(x^i, x^n)\right\} \quad (2)$$

.

What' more, we make use of multi-scale saliency detection in order to improve the contrast between salient and non-salient region. Therefore, we compute S^i using four scales $\{100\%, 80\%, 50\%, 30\%\}$, then compute the average value $\overline{S^i}$ of them.

2.2. Category-specific Salient Region Filtering

After achieving salient patches with the highest saliency score from each scene image, some patches we got are not special and have few information which can not be used for scene classification. Therefore our goal is to get the category-specific salient region(CSSR) which can reveal the essential features and to learn CSSR filters per category.

Image Feature Descriptor We represent each $n \times n$ patch with multi-scale DERF[5] feature. DERF is proposed by our research group, which is a new distinctive efficient robust descriptor, inspired by modeling of the response and distribution properties of the P ganglion cells. DERF convolve gradient maps at the locations of grid points using the DoG function, and the grid points are arranged into concentric rings which its radial distance increase exponentially.

CSSR modeling and filtering The salient image regions, which have high frequency occurring in the identical scene category but distinguish from the other scene categories, are qualified to be chosen as the category-specific salient region(CSSR). Following the method in the [7], we first do clustering per category for salient regions we got, then learn CSSR filters based on the cluster result and CSSR's appearance and saliency. At last, we adopt CSSR filters convoluting with input images to achieve CSSR and construct the CSSR representation.

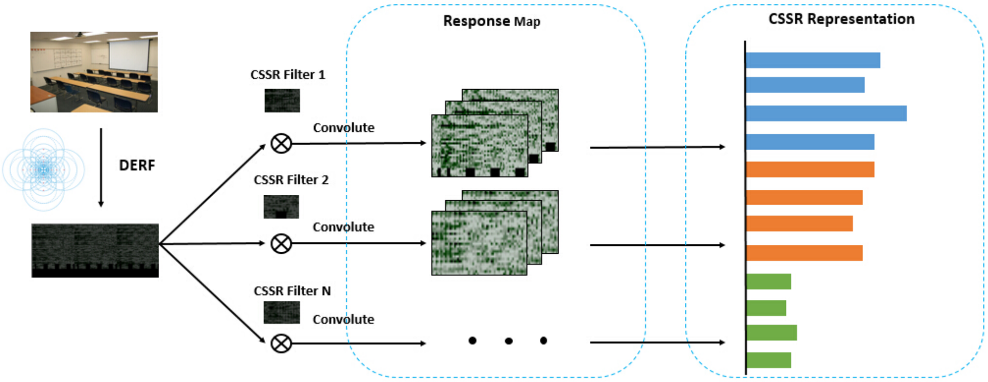


Fig. 2. Construction of CSSR representation for scene classification.

Through clustering the salient regions we got, the CSSR that belong to the same category always locate within a few clusters, such as regions consist of desks, chairs and blackboard in classroom. It is obvious that these regions have high possibility to be found within clusters, and the common regions occurring in many categories often locate far away from the clusters. So we model the relationship between images and spatial clusters. We define the potential of CSSR in the pixel level occurring in location p as

$$\phi(p, p^c, \sigma) = exp(-\frac{\| p - p^c \|^2}{\sigma}) \tag{3}$$

where p and p^c are position of the salient region and its cluster center, σ is the parameter controlling the coverage of CSSR. So the union process using a mixture model

$$\Phi(p, p^c, \sigma) = \sum_{i=1}^{k} d_i \cdot \phi_d(p, p_i^c, \sigma_i) \tag{4}$$

where k is the number of cluster and d_i is the weight of each CSSR.

Because the scene can be classified using several CSSRs, we define the multi-region joint model of CSSR's occurrence at position p_i in image I is

$$f(I,P) = \sum_{i=0}^{n} F_i \cdot H(I,p_i) + \sum_{i=1}^{n}\sum_{j=1}^{k} d_{ij} \cdot \Phi(p_i, p_{ij}^c, \sigma_{ij}) + S^{p_i} \tag{5}$$

where the $F_i \cdot H(I, p_i)$ is the appearance term in convolution to achieve the response map, F_i is the CSSR filter vector extracted from position p_i in image I, $H(I, p_i)$ is the feature vector using DERF and S^{p_i} is salient value in position p of image I. Also, $P = [p_0, ..., p_n]$, n and k are the numbers of salient regions and the center corresponding to the i^{th} salient region. p_{ij}^c is the central location of the j^{th} cluster corresponding to the i^{th} part.

2.3. Scene Image Representation and Classification

Using the inference and learning algorithm in [7], we can get the CSSR filters and parameters. In feature pooling, we extract value $A = F_i \cdot H(I, p_i)$ and $S = \sum_{j=1}^{k} d_{ij} \cdot \phi(p_i, p_{ij}^c, \sigma_{ij}) + S^{p_i}$ from image I. Then we concatenate all the set of values A and S into a vector as the CSSR representation of the image. Suppose we have N salient region filters and perform feature pooling on R scale, then then dimensionality of the CSSR representation is $2NR$. The process of construct CSSR representation using CSSR filters are shown in Figure 2.

Also we extract whole image deep feature through Places CNN[8], and concatenate with CSSR representation to be the Deep-CSSR feature. Then training a neural network to do scene images classification, which has two fully-connected hidden layers, 200 nodes and adopt rectified linear function(ReLU) as the activation function.

3. EXPERIMENTS

In this section, we conducted experiments to demonstrate the effectiveness of our framework with two databases: MIT-Indoor 67[9] and Sun397[10] databases.

3.1. Database and experiment setup

MIT Indoor 67 database It include 15,620 indoor scene images in 67 scene classes, and we select 80 images for training and 20 images for testing each scene category. These indoor images has so complicate layouts and different objects that it is challenging for scene image classification.

SUN397 database It contains 397 scene categories and 108,754 images in sum. It is a large-scale scene database which is suitable for deep CNN learning. In our experiment, we adopt 50 images for training and 50 images for testing from every scene class followed the traditional evaluation

Method	Accuracy(%)
SPM[3]	34.40
OTC[11]	47.33
Discriminative Patches ++[12]	49.40
ImageNet-CNN[8]	56.79
SPM+OPM[13]	63.48
Mid-level Elements[14]	66.87
ISPR+IFV[7]	68.50
Places-CNN[8]	68.24
MOP-CNN[15]	68.88
Hybrid-CNN[8]	70.80
ImgNET fc8-FV[16]	72.86
DSFL+DeCAF[17]	76.23
DEEP-CSSR	**77.80**

Table 1. Scene Classification Performance on MIT Indoor 67

Method	Accuracy(%)
OTC[11]	34.56
DeCAF[18]	40.94
ImageNet-CNN[8]	42.61
SPM+OPM[13]	45.91
FV[19]	47.20
MTL-SDCA,Sqr[20]	49.50
OTC+HOG2×2[11]	49.60
MOP-CNN[15]	51.98
Hybrid-CNN[8]	53.86
Places-CNN[8]	54.32
ImgNET fc8-FV[16]	54.40
DEEP-CSSR	**57.30**

Table 2. Scene Classification Performance on SUN397

benchmark.

Based on the result of salient regions clustering, we respectively choose 6 and 8 top ranked cluster centers from training images per category on MIT indoor 67 and SUN397 databases. So we apply 6 and 8 category-specific salient regions(CSSR) filters per category and perform on 4 scales for each image on two databases. It is demonstrated that these parameters chosen flexibly based on the database could get the better performance than the methods which set fixed parameters. Therefore, in MIT indoor 67 every category has 6 CSSR filters, and the total number of CSSR filters is $67 \times 6 = 402$, and we can get the $2 \times 402 \times 4 = 3216$ dimensions CSSR representation for each image. In SUN397 the total number of CSSR filters is $397 \times 8 = 3176$, and we can get the $2 \times 3176 \times 4 = 25408$ dimensions CSSR representation for each image. Then, we get the whole scene image representation by concatenated feature vector of CSSR representation and Hybrid CNN feature of the whole image that have 4096 dimensional vector and 7 layers. We do scene image classification by a neural network that have two fully-connected layers and 200 nodes each layer with the whole image representation.

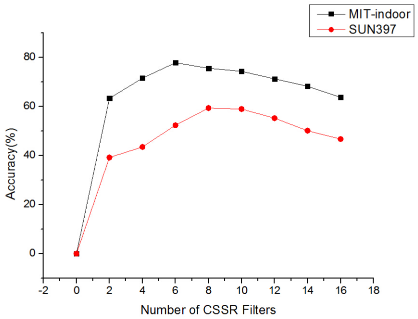


Fig. 3. Classification accuracy vs. number of CSSR filters on datasets

3.2. Results and analysis

We compare the classification accuracy of our approach with other results published on MIT Indoor 67 database and SUN397 database. These existing methods consist of methods that apply effective feature descriptors, such as oriented texture curves(OTC)[11], spatial pyramid matching(SPM)[3], and mid-level feature methds[12]. Also, taking deep learning methods into consideration, conclude MOP-CNN[15], Place-CNN[8] and Hybrid-CNN[8]. As can be seen from Table 1 and Table 2, higher classification accuracy can be achieved by our novel pipeline.

Also, we perform an study to analyze the impact of the number of CSSR filters per category in our framework. The different classification accuracy results on MIT-indoor and SUN397 are shown in Figure 3. This controlled experiment reveals that the CSSR filters are important in our method, and too few and too many CSSR filters would lead to poor performance.

4. CONCLUSION AND FUTURE WORK

In this paper we have presented a novel framework that utilizes category-specific salient region with deep CNN feature(DEEP-CSSR) for improving scene classification. We adopt an unsupervised salient region detection method, an inference and learning algorithm to achieve the category-specific salient region filters, and construct the DEEP-CSSR features to do classification. We have evaluated our method on MIT indoor67 and SUN397 datasets, and obtained the better results than other resent works. It is demonstrated that saliency detection combine with deep learning for visual recognition and classification task will be promising.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Hong Kong, Macao, and Taiwan Science and Technology Cooperation Program of China(Grant No.L2015TGA9004).

6. REFERENCES

[1] David G Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Engin Tola, Vincent Lepetit, and Pascal Fua, “Daisy: An efficient dense descriptor applied to wide-baseline stereo,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 815–830, 2010.

[3] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Conference on*. 2006, vol. 2, pp. 2169–2178, IEEE.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] Dawei Weng, Yunhong Wang, Mingming Gong, Dacheng Tao, Hui Wei, and Di Huang, “Derf: Distinctive efficient robust features from the biological modeling of the p ganglion cells,” *Image Processing, IEEE Transactions on*, vol. 24, no. 8, pp. 2287–2302, 2015.

[6] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, “Context-aware saliency detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 1915–1926, 2012.

[7] Di Lin, Cewu Lu, Renjie Liao, and Jiaya Jia, “Learning important spatial pooling regions for scene classification,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. 2014, pp. 3726–3733, IEEE.

[8] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, “Learning deep features for scene recognition using places database,” in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[9] Ariadna Quattoni and Antonio Torralba, “Recognizing indoor scenes,” in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. 2009, pp. 413–420, IEEE.

[10] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva, “Sun database: Exploring a large collection of scene categories,” *International Journal of Computer Vision*, pp. 1–20, 2014.

[11] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal, *OTC: A Novel Local Descriptor for Scene Classification*, pp. 377–391, Springer, 2014.

[12] Saurabh Singh, Abhinav Gupta, and Alexei Efros, “Unsupervised discovery of mid-level discriminative patches,” *Computer VisionCECCV 2012*, pp. 73–86, 2012.

[13] Lingxi Xie, Jingdong Wang, Baining Guo, Bo Zhang, and Qi Tian, “Orientational pyramid matching for recognizing indoor scenes,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. 2014, pp. 3734–3741, IEEE.

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros, “Mid-level visual element discovery as discriminative mode seeking,” in *Advances in Neural Information Processing Systems*, 2013, pp. 494–502.

[15] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, *Multi-scale orderless pooling of deep convolutional activation features*, pp. 392–407, Springer, 2014.

[16] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, Nuno Vasconcelos, Weixin Li, and Nuno Vasconcelos, “Scene classification with semantic fisher vectors,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 2974–2983.

[17] Zhen Zuo, Gang Wang, Bing Shuai, Lifan Zhao, Qingxiong Yang, and Xudong Jiang, *Learning discriminative and shareable features for scene classification*, pp. 552–568, Springer, 2014.

[18] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International Conference on Machine Learning*, 2013, pp. 647–655.

[19] Jorge Snchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, “Image classification with the fisher vector: Theory and practice,” *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[20] Maksim Lapin, Bernt Schiele, and Matthias Hein, “Scalable multitask representation learning for scene classification,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. 2014, pp. 1434–1441, IEEE.