Better Exploiting OS-CNNs for Better Event Recognition in Images

Limin Wang Zhe Wang Sheng Guo Yu Qiao
Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS, China
{07wanglimin, buptwangzhe2012, guosheng1001}@gmail.com, yu.qiao@siat.ac.cn

Abstract

Event recognition from still images is one of the most important problems for image understanding. However, compared with object recognition and scene recognition, event recognition has received much less research attention in computer vision community. This paper addresses the problem of cultural event recognition in still images and focuses on applying deep learning methods on this problem. In particular, we utilize the successful architecture of Object-Scene Convolutional Neural Networks *(OS-CNNs) to perform event recognition. OS-CNNs are composed of object nets and scene nets, which transfer the learned representations from the pre-trained models on large-scale object and scene recognition datasets, respectively. We propose four types of scenarios to explore OS-CNNs for event recognition by treating them as either "end-to-end event predictors" or "generic feature extractors". Our experimental results demonstrate that the global and local representations of OS-CNNs are complementary to each other. Finally, based on our investigation of OS-CNNs, we come up with a solution for the cultural event recognition track at the ICCV ChaLearn Looking at People (LAP) challenge 2015. Our team secures the third place at this challenge and our result is very close to the best performance.*

1. Introduction

Image understanding [12, 18, 20, 27] is becoming one of the most important problems in computer vision and many research efforts have been devoted to this topic. While object recognition [4] and scene recognition [28] have been extensively studied in the task of image classification, event recognition [14, 23, 26] in still images received much less research attention, which also plays an important role in semantic image interpretation. As shown in Figure 1, the characterization of event is extremely complicated as the event concept is highly related to many other high-level visual cues, such as objects, scene categories, human garments, human poses, and other context. Therefore, event recognition in still images poses more challenges for the current state-of-the-art image classification methods, and needs to be further investigated in the computer vision research.

Figure 1. Examples of cultural event images from the ICCV ChaLearn Looking at People (LAP) dataset. From these examples, we can see that the characterization of event is complicated and it is related to many visual cues, such as objects, secne category, and human garments.

Convolutional neural networks (CNNs) [13] have recently enjoyed great successes in large-scale image classification, in particular for object recognition [9, 18, 20] and scene recognition [21, 28]. For event recognition, much fewer deep learning methods have been designed for this problem. Our previous work [23] proposed a new deep architecture, called *Object-Scene Convolutional Neural Network* (OS-CNN), for cultural event recognition. OS-CNNs are designed to extract useful information for event understanding from the perspectives of containing objects and scene categories, respectively. OS-CNNs are composed of two-stream CNNs, namely object nets and scene nets. Object nets are pre-trained on the large-scale object recognition datasets (e.g. ImageNet [4]), and scene nets are based on models learned from the large-scale scene recognition datasets (e.g. Places205 [28]). Decomposing into object nets and scene nets enables us to use the external large-scale annotated images to initialize OS-CNNs, which may be further fine tuned elaborately on the event recognition dataset. Finally, event recognition is performed based on the late fusion of softmax outputs of object nets and scene nets.

Following the research line of OS-CNNs, in this paper, we try to further explore different aspects of OS-CNNs

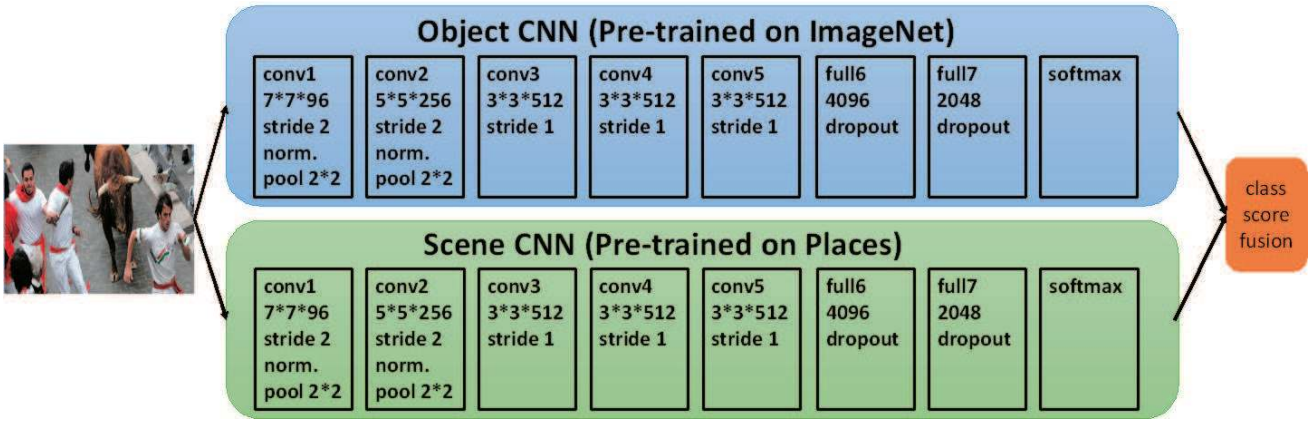

Figure 2. The architecture of Object-Scene Convolutional Neural Network (OS-CNN) for event recognition from [23]. OS-CNN is composed to two-stream networks: object nets and scene nets, which are separately pre-trained on the ImageNet and Places205 dataset.

and better exploit OS-CNNs for better event recognition. Specifically, we design four types of investigation scenarios to study the performance of OS-CNNs. In the first scenario, we directly use the softmax outputs of CNNs as recognition results. In the next three scenarios, we treat CNNs as feature extractors, and use them to extract both *global* and *local* features of an image region. Global features are more compact and aim to capture the holistic structure, while local features focus on describing the image details and local patterns. Our experimental results indicate these two kinds of features are complementary to each other and robust for event recognition. Based on our empirical explorations with OS-CNNs, we come up with our solution for the cultural event recognition track at the ICCV ChaLearn Looking at People (LAP) challenge [6] and we secure the third place.

The rest of this paper is organized as follows. In Section 2, we will give a brief introduction to OS-CNNs, including network architectures and implementation details. After that, we will introduce our extensive explorations with OS-CNNs for event recognition in Section 3. We then report our experimental results in Section 4. Finally, we conclude our method and present the future work in Section 5.

2. OS-CNNs Revisited

In this section, we will first briefly introduce the architecture of *Object-Scene Convolutional Neural Networks* (OS-CNNs), which was proposed in our previous work [23]. Then, we will present the implementation details of OS-CNNs, including network structures, data augmentations, and learning policy.

2.1. OS-CNNs

Event is a relatively complicated concept in computer vision research and highly related with other two problems: object recognition and scene recognition. The basic idea behind OS-CNN is to utilize two separate components to perform event recognition from the perspectives of occurring objects and scene context. Specifically, OS-CNNs are composed of object nets and scene nets, as shown in Figure 2 .

Object nets. Object net is designed to capture useful information of objects to help event recognition. Intuitively the occurring objects are able to provide useful cues for event understanding. For instance, in the cultural event of Australia Day as shown in Figure 1, Australian flag will be a representative object. As the main goal of object net is to deal with object cues, we build it based on recent advances on large-scale object recognition, and pre-train the network on the public ImageNet models. Then, we further fine tune the model parameters on the training dataset of cultural event recognition by setting the output number as 100 (cultural event recognition dataset containing 100 classes).

Scene nets. Scene net is expected to extract scene information of image to assist event understanding. In general, the scene context will be helpful for recognizing the event category in the image. For example, in the cultural event of Sapporo Snow Festival as shown in Figure 1, outdoor will be usually the scene category. Specifically, we pre-train the scene nets by using the models learned on the dataset Places205, which contains 205 scene classes and 2.5 millions images. Similar to object nets, we then fine tune the network weights of scene nets on the event recognition dataset, where we set network output number as 100.

Based on the above analysis, recognizing cultural event will benefit from the transferred representations learned for object recognition and scene recognition. Thus, we will fuse the network outputs of both object nets and scene nets as the prediction of OS-CNNs.

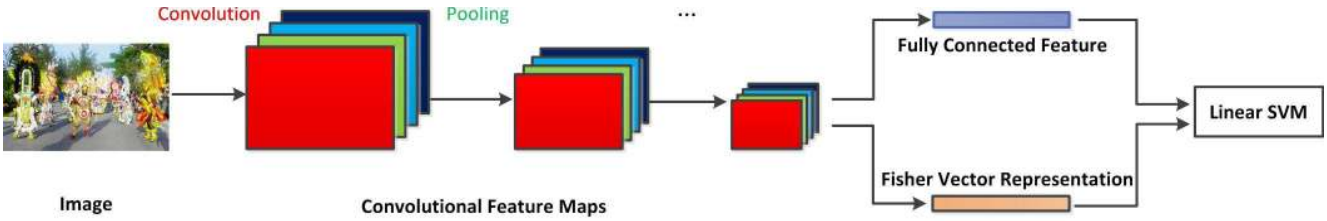

Figure 3. Our better explorations with OS-CNNs for event recognition. We utilize OS-CNNs to extract both global representations (activations of fully connected layers) and local representations (activations of convolutional layers), which can be combined for event recognition in still images.

2.2. Implementation details

In this subsection, we will describe the implementation details of training OS-CNNs, including network structures, data augmentations, and learning policy.

Network structures. Network structures are of great importance for improving the performance of CNNs. In the past several years, many successful network architectures have been proposed for object recognition, such as AlexNet [12], ClarifaiNet [27], OverFeat [17], GoogLeNet [20], VGGNet [18], MSRANet [9], and Inception2 [10]. Some good practices can be drawn from the evolution of network architectures: smaller convolutional kernel size, smaller convolutional stride, more convolutional channel, deeper network structure. In this paper, we choose the VGGNet-19 as our main investigated structure due to its good performance in object recognition, which is composed of 16 convolutional layers and 3 fully connected layers. The detailed description about VGGNet-19 is out of the scope of this paper and can be found in [18].

Data augmentations. By data augmentation, we mean perturbing an image by transformations that leave the underlying class unchanged. Typical transformations include corner cropping, scale jittering, and horizontal flipping. Specifically, during the training phase of OS-CNNs, we randomly crop image regions (224×224) from 4 corners and 1 center of the whole image. Meanwhile these cropped regions undergo horizontal flipping randomly. Furthermore, we use three different scales to resize training images, where the smallest size s of an image is set to $256, 384, 512$.

It should be noted that data augmentation is a method applicable to both training images and testing images. During training phase, data augmentation will generate additional training examples and reduce the influence of overfitting. For testing phase, data augmentation will help to improve the classification accuracy. The augmented samples can be either regarded as independent images or combined into a single representation by pooling or stacking operations. In the current implementation, during the test phase, we use sum pooling to aggregate these representations of augmented samples into a single representation.

Learning policy. Effective training methods are very crucial for learning CNN models. As the training dataset of cultural event recognition is relatively small compared with ImageNet [4] and Places205 [28], we resort to pre-training OS-CNNs by using these public available models trained on ImageNet and Places205. Specifically, we pre-train object nets with public VGGNet-19 model [1], which achieved the top performance at ILSVRC2014. For scene net, we use the model released by [21] [2] to initialize the network weights, which has obtained the best performance on the Places205 dataset so far.

The network weights are learned using the mini-batch stochastic gradient descent with momentum (set to 0.9). At each iteration, a mini-batch of 256 images is constructed by random sampling. The dropout ratios for fully connected layers are set as 0.5. As we pre-train network weights with ImageNet and Places205 models, we set a smaller learning rate for fine tuning OS-CNNs: learning rate starts with 10^{-3}, decreases to 10^{-4} after 5K iterations, decreases to 10^{-5} after 10K iterations and the training process ends at 12K iterations. To speed up the training process, we use a Multi-GPU extension version [24] of Caffe toolbox [11], which is publicly available online [3].

3. Exploring OS-CNNs

We have introduced the architectures and implementation details about OS-CNNs in the previous section. In this section, as shown in Figure 3, we will focus on describing the explorations of OS-CNN activations from different layers and try to improve the recognition performance.

3.1. Scenario 1: OS-CNN predictions

The simplest way to utilize OS-CNNs for cultural event recognition is directly using the outputs (softmax layer) of CNN networks as final prediction results. Specifically, given an image $\mathbf{I}$, its recognition score is calculated as follows:

$$s_{os}(\mathbf{I}) = \alpha_o s_o(\mathbf{I}) + \alpha_s s_s(\mathbf{I}), \quad (1)$$

[1] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

[2] https://github.com/wanglimin/Places205-VGGNet

[3] https://github.com/yjxiong/caffe

where $s_o(\mathbf{I})$ and $s_s(\mathbf{I})$ are the prediction scores of object nets and scene nets, α_o and α_s are the fusion weights of object nets and scene nets. In the current implementation, fusion weights are set to be equal for object nets and scene nets.

3.2. Scenario 2: OS-CNN global representations with pre-training

Another way to deploy OS-CNNs for cultural event recognition is to treat them as generic feature extractors and use them to extract the global representation of an image region. We usually extract the activations of **fully connected layers**, which are very compact and discriminative. In this case, we only use the pre-trained models without fine-tuning. Specifically, given an image region I, we extract this global representation based on OS-CNNs as follows:

$$\phi^p_{os}(\mathbf{I}) = [\beta_o \phi^p_o(\mathbf{I}), \ \ \beta_s \phi^p_s(\mathbf{I})], \tag{2}$$

where $\phi^p_o(\mathbf{I})$ and $\phi^p_s(\mathbf{I})$ are the CNN activations from pre-trained object nets and scene nets, β_o and β_s are the fusion weights of object nets and scene nets. In current implementation, the fusion weights are set to be equal for object nets and scene nets.

3.3. Scenario 3: OS-CNN global representations with pre-training and fine-tuning

In previous scenario, OS-CNNs are only pre-trained on large scale dataset of object recognition and scene recognition, and directly applied to the smaller event recognition dataset. However, it was demonstrated that fine-tuning a pre-trained CNNs on the target data can improve the performance a lot [8]. We consider fine-tuning the OS-CNNs on the event recognition dataset and the resulted image representations become dataset-specific. After fine-tuning process, we obtain the following global representation with the fine-tuned OS-CNNs:

$$\phi^f_{os}(\mathbf{I}) = [\beta_o \phi^f_o(\mathbf{I}), \ \ \beta_s \phi^f_s(\mathbf{I})], \tag{3}$$

where $\phi^f_o(\mathbf{I})$ and $\phi^f_s(\mathbf{I})$ are the CNN activations from the fine-tuned object nets and scene nets, β_o and β_s are the fusion weights of object nets and scene nets. In current implementation, the fusion weights are set to be equal for object nets and scene nets.

3.4. Scenario 4: OS-CNN local representations + Fisher vector

In previous two scenarios, we extract a global representation of an image region with OS-CNNs. Although this global representation is compact and discriminative, it may lack the ability of describing local patterns and detailed information. Inspired by the recent success on video-based action recognition with deep convolutional descriptors [22], we investigate the effectiveness of **convolutional layer** activations. Convolutional layer features have been also demonstrated to be effective in image-based tasks, such as object recognition [7], scene recognition [5] and texture recognition [3]. In this scenario, OS-CNNs are first pre-trained on large-scale ImageNet and Places205 datasets, and then fine-tuned on the event recognition dataset, just as in Scenario 3.

Specifically, given an image region $\mathbf{I}$, we first extract the convolutional feature maps of OS-CNNs (activations of convolutional layers) $C(\mathbf{I}) \in \mathbb{R}^{n \times n \times c}$, where n is feature map size and c is feature channel number. Each activation value in the convolutional feature map corresponds to a local receptive field in the original image, and therefore we call these activations of convolutional layers as OS-CNN local representations.

After extracting OS-CNN local representations, we utilize two normalization methods, namely *channel normalization* and *spatial normalization* proposed in [22], to preprocess these convolutional feature maps into transformed convolutional feature maps $\widetilde{C}(\mathbf{I}) \in \mathbb{R}^{n \times n \times c}$. More details regarding these two normalization methods are out scope of this paper and can be found in [22]. The normalized CNN activation $\widetilde{C}(\mathbf{I})(x, y, :) \in \mathbb{R}^c$ at each postion (x, y) is called as the *Transformed Deep-convolutional Descriptor* (TDD). These two kinds of normalization methods have turned out to be effective for improving the performance of CNN local representations in [22]. Moreover, the combination of them can obtain higher performance. Therefore, we will use both normalization methods in our experimental explorations.

Finally, we employ Fisher vector [16] to encode these TDDs into a global representation due to its good performance in object recognition [2] and action recognition [19, 25]. In particular, according to our previous comprehensive study on encoding methods [15], we first use PCA to reduce the dimension of TDD to 64. Then each TDD is soft-quantized with a Gaussian Mixture Model (GMM) with K components (K set to 256). The first and second order differences between each TDD $\mathbf{x} \in \mathbb{R}^{64}$ and its Gaussian center μ_k are aggregated in the block $\mathbf{u}_k$ and $\mathbf{v}_k$, respectively. The final Fisher vector representation is yielded by concatenating these blocks together:

$$\phi_{fv}(\mathbf{I}) = [\mathbf{u}_1, \mathbf{v}_1, \cdots, \mathbf{u}_K, \mathbf{v}_K]. \tag{4}$$

For OS-CNNs, the Fisher vector of local representation is defined as follows:

$$\phi^f_{os-fv}(\mathbf{I}) = [\beta_o \phi^f_{o-fv}(\mathbf{I}), \ \ \beta_s \phi^f_{s-fv}(\mathbf{I})], \tag{5}$$

where $\phi^f_{o-fv}(\mathbf{I})$ is the Fisher vector representation from object nets, $\phi^f_{s-fv}(\mathbf{I})$ is the Fisher vector representation from scene nets, β_o and β_s are their fusion weights and set to be equal to each other in the current implementation.

	Object nets	Scene nets	OS-CNNs
Scenario 1			
softmax	73.1%	71.2%	75.6%
Scenario 2			
fc7	67.2%	63.4%	69.1%
Scenario 3			
fc6	80.6%	76.8%	81.7%
fc7	81.4%	78.1%	82.3%
Scenario 4			
conv5-1	77.6%	76.6%	78.9%
conv5-2	78.6%	76.2%	79.6%
conv5-3	79.4%	76.1%	80.2%
conv5-4	78.4%	75.6%	79.7%
Fusion			
conv5-3+fc7	82.5%	79.3%	83.2%

Table 1. Event recognition performance of OS-CNN global and local representations on the validation data.

3.5. Linear classifiers

All the representations $\phi(\mathbf{I})$ in previous three scenarios are used to construct a linear classifier $s(\mathbf{w}, \mathbf{I}) = \mathbf{w}\phi(\mathbf{I})$, where $\mathbf{w}$ is the weight of linear classifier. In our implementation, we choose LIBSVM [1] as the classifier to learn the weight $\mathbf{w}$, where the parameter C, balancing regularizer and loss, is set as 1. It is worth noting that all these representations are first normalized before fed into SVM for training. For OS-CNN global representations, we use ℓ_2-normalization, and for OS-CNN local representations, we use intra normalization and power ℓ_2-normalization.

4. Experiments

In this section, we first describe the dataset of cultural event recognition at the ICCV ChaLearn Looking at People (LAP) challenge 2015. Then we present and analyze the experimental results of our proposed different representations with OS-CNNs on the validation dataset of ChaLearn LAP dataset. Finally, we describe our solution for the ICCV ChaLearn LAP challenge 2015.

4.1. Datasets and evaluation protocol

Datasets. The ICCV ChaLearn LAP challenge 2015 [6] contains a track of cultural event recognition and provides an event recognition dataset. This dataset contains images collected from two image search engines (Google Images and Bing Images). There are totally 100 event classes (99 event classes and 1 background class) from different countries and some images are shown in Figure 1. From these samples, we see that cultural event recognition is really complicated, where garments, human poses, objects and scene context all constitute the possible cues to be exploited for event understanding. This dataset is divided into three parts: development data (14,332 images), validation data (5,704 images), and evaluation data (8,669 images). As we can not access the label of evaluation data, we mainly train our models on the development data and report the results on the validation data.

Rank	Team	Score
1	VIPL-ICT-CAS	85.4%
2	FV	85.1%
3	**MMLAB (ours)**	84.7%
4	NU&C	82.4%
5	CVL_ETHZ	79.8%
6	SSTK	77.0%
7	MIPAL_SUN	76.3%
8	ESB	75.8%
9	Sungbin Choi	62.4%
10	UPC-STP	58.8%

Table 2. Comparison the performance of our submission with those of other teams. Our team secures the third place in the ICCV ChaLearn LAP challenge 2015.

Evaluation protocol. The principal quantitative measure is based on precision recall curve. They use the area under this curve as the computation of the average precision (AP), which is calculated by numerical integration. Finally, they average these per-class AP values across all event classes and employ the mean average precision (mAP) as the final ranking criteria. Hence, in our exploration experiments, we report our results evaluated as AP value for each class and mAP value for all classes.

4.2. Results and analysis

Settings. In this exploration experiment, we use the VGGNet-19 as the OS-CNN network structure. We extract activations from two fully connected layers (`fc6`, `fc7`) as OS-CNN global representations, and activations from four convolutional layers (`conv5-1`, `conv5-2`, `conv5-3`, `conv5-4`) as OS-CNN local representations. It should be noted that we choose the activations after rectified Linear Units (ReLUs). We use ℓ_2-normalization to further process OS-CNN global representations for better SVM training. For Fisher vector representation of OS-CNN local representation, we employ intra-normalization and power ℓ_2-normalization, as suggested by [15].

Analysis. We first report the numerical results in Table 1. From these results, several conclusions can be drawn as follows:

- We see that the object nets outperform scene nets on the task of cultural event recognition, which may imply that object cues play more important roles than scene cues for cultural event understanding.

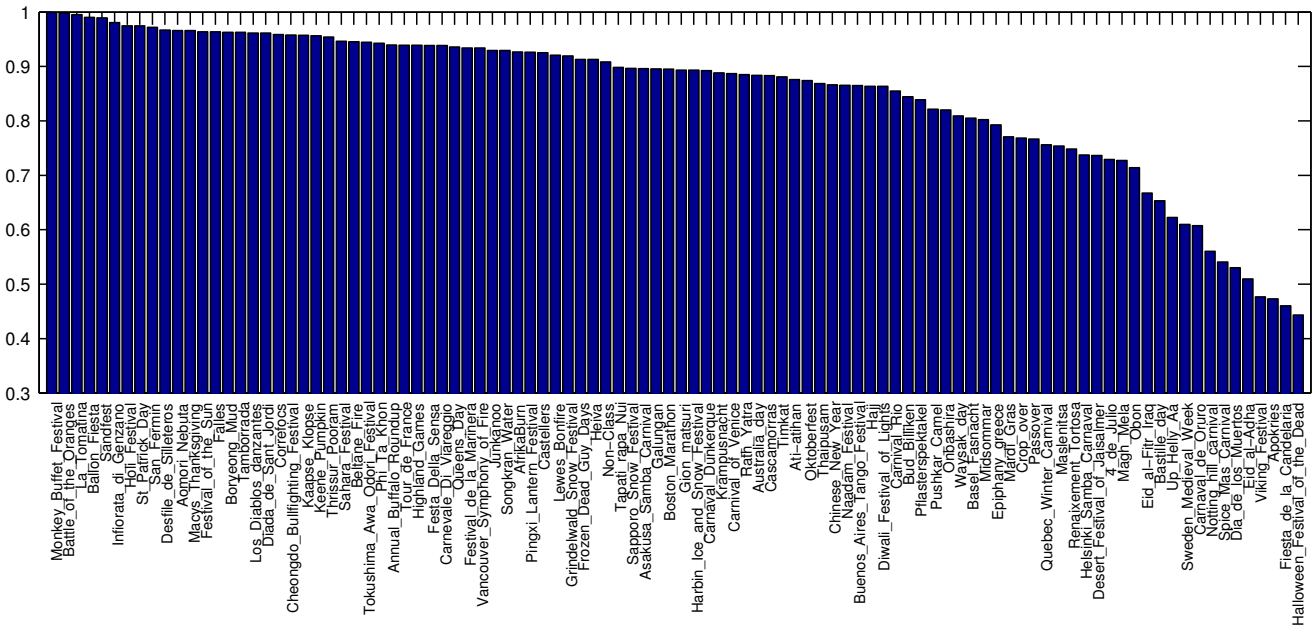

Figure 4. Per-class AP value of combining OS-CNN global and local representations on the validation data of ICCV ChaLearn LAP dataset.

- We observe that OS-CNNs are effective for event recognition as it extract both object and scene information from the image. They achieve superior performance to object nets and scene nets, no matter what scenario is adopted.

- We can notice that combining fine tuned features with linear SVM classifier (scenario 3) is able to obtain better performance than direct using the softmax output of CNNs (scenario 1). This result may be ascribed to the fact that CNNs are easily over-fitted to the training samples when the number of training images is relatively small.

- Comparing fine-tuned features (scenario 3) with pre-trained features (scenario 2), we may conclude that fine tuning on the target dataset is very useful for improving recognition performance, which agrees with the findings of [8].

- Comparing the local representations (scenario 4) and global representations (scenario 3) of CNNs, we see that global representation achieve slightly higher recognition accuracy.

- We further combine the global representation (`fc7`) with local representation (`conv5-3`) of CNNs and find that this combination is capable of boosting final recognition performance. This performance improvement indicates that different layers of CNNs capture different level abstraction of original image. These feature activations from different layers are complementary to each other.

We also plot the AP values for all event classes in Figure 4. From these AP values, we see that the events of `Monkey Buffet Festival` and `Battle of the Oranges` achieve the highest performance (100%). This result may be ascribed to the fact that there are specific objects in these two event categories. At the same time, we notice that some event classes obtain very low AP values, such as `Halloween Festival of the Dead`, `Fiesta de la Candelaria`, `Apokries`, and `Viking Festival`. The AP values of these cultural event classes are below 50%. In general, there are no specific objects and scene context in these difficult event classes, and besides these classes are easily confused with other classes from the perspective of visual appearance, as observed from Figure 5.

We visualize several recognition examples in Figure 5. In the row 1 we give eight examples that are successfully predicted by our method, from classes like `Keene Pummpking`, `Boryeong Mud`, `AfrikaBurn` and so on. Meanwhile, we also provide some failure cases with high confidence from our method in the rows 2,3,4. From these wrong predicted examples, we see that these failure cases are rather reasonable and there exists great confusion between some cultural event classes. For example, the event classes of `Dia de los Muertos` and `Halloween Festival of the Dead` share similar human make-up and garments. The event classes of `Up Helly Aa` and `Viking Festtival` share similar hu-

Figure 5. Examples of images that our method succeeds and fails in top-1 evaluation. We give 8 successfully predicted and 24 wrong predicted images in the row 1 and rows 2,3,4, respectively.

man dresses and containing objects. The event classes of `Harbin Icen and Snow Festival` and `Sapporo Snow Festival` share similar scene context and color appearance. The event classes of `Chinese New Year` and `Pingxi Lattern Festival` share similar containing objects. In summary, these examples in Figure 5 indicate that the concept of event is really complicated and there only exist slight difference between some event classes.

4.3. Challenge results

For final evaluation, we merge the development data (14,332 images) and validation data (5,704 images) into a single training dataset (20,036 images) and re-train our OS-CNN models on this new dataset. Our final submission results to the ICCV ChaLearn LAP challenge are based on our re-trained model.

According to the above experimental explorations, we conclude that the OS-CNN global and local representations are complementary to each other. Thus, we choose to combine activations from `fc7` and `conv5-3` layers, to keep a balance between performance and efficiency. Meanwhile, our previous study demonstrated that GoogLeNet is complementary to VGGNet [23]. Hence, we also extract a global representation by using the OS-CNNs of GoogLeNet in our challenge solution. In summary, our challenge solution is composed of three representations: (i) OS-CNN VGGNet-19 local representations, (ii) OS-CNN VGGNet-19 global representations, and (iii) OS-CNN GoogLeNet global representations.

The challenge results are summarized in Table 2. We see that our method is among the top performers and our mAP is very close to the best performance of this challenge (84.7% vs. 85.4%). Regarding computational cost, our implementation is based on CUDA 7.0 and Matlab 2013a, and it takes about 1s to process one image in our workstation equipped with 8 cores CPU, 48G RAM, and Tesla K40 GPU.

5. Conclusions

In this paper, we have comprehensively studied different aspects of OS-CNNs for better cultural event recognition. Specifically, we investigate the effectiveness of CNN activations from different layers by designing four types scenarios of adapting OS-CNNs to the task of cultural event recognition. From our empirical study, we demonstrate that the CNN activations from convolutional layers and fully connected layers are complementary to each other, and the combination of them is able to boost recognition performance. Finally, we come up with a solution by using OS-CNNs at the ICCV ChaLearn LAP challenge and secure the third place. In the future, we may consider how to incorporate more visual cues such as human poses, garments, object and scene relationship in a systematic manner for event recognition in still images.

Acknowledgement

This work is supported by a donation of two Tesla K40 GPUs from NVIDIA Corporation. Meanwhile this work is partially supported by National Natural Science Foundation of China (91320101, 61472410), Shenzhen Basic Research Program (JCYJ20120903092050890, JCYJ20120617114614438, JCYJ20130402113127496), 100 Talents Program of CAS, and Guangdong Innovative Research Team Program (No.201001D0104648280).

References

[1] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):27, 2011. 5

[2] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, pages 1–12, 2011. 4

[3] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *CoRR*, abs/1507.02620, 2015. 4

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 3

[5] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *CVPR*, pages 2974–2983, 2015. 4

[6] S. Escalers, J. Fabian, P. Pardo, X. Baro, J. Gonzlez, H. J. Escalante, and I. G. and. Chalearn 2015 apparent age and cultural event recognition: datasets and results. In *ICCV, ChaLearn Looking at People workshop*, 2015. 2, 5

[7] B. Gao, X. Wei, J. Wu, and W. Lin. Deep spatial pyramid: The devil is once again in the details. *CoRR*, abs/1504.05277, 2015. 4

[8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 4, 6

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. 1, 3

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 3

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014. 3

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012. 1, 3

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 1

[14] L. Li and F. Li. What, where and who? classifying events by scene and object recognition. In *ICCV*, pages 1–8, 2007. 1

[15] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CoRR*, abs/1405.4506, 2014. 4, 5

[16] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 4

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 3

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 3

[19] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In *WACV*, pages 15–22, 2013. 4

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 1, 3

[21] L. Wang, S. Guo, W. Huang, and Y. Qiao. Places205-VGGNet models for scene recognition. *CoRR*, abs/1508.01667, 2015. 1, 3

[22] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 4

[23] L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. In *CVPR, ChaLearn Looking at People 2015 workshop*, pages 30–35, 2015. 1, 2, 7

[24] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *CoRR*, abs/1507.02159, 2015. 3

[25] X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *ACCV*, pages 572–585, 2012. 4

[26] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *CVPR*, pages 1600–1609, 2015. 1

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 1, 3

[28] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014. 1, 3