2015 IEEE International Conference on Computer Vision Workshops

Exploiting Feature Hierarchies with Convolutional Neural Networks for Cultural Event Recognition

Mengyi Liu*[1], Xin Liu*[1], Yan Li[1], Xilin Chen[1], Alexander G. Hauptmann[2],Shiguang Shan[1]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]School of Computer Science, Carnegie Mellon University, 15213, USA
{mengyi.liu, xin.liu, yan.li}@vipl.ict.ac.cn, {xlchen, sgshan}@ict.ac.cn, alex@cs.cmu.edu

Abstract

Cultural events are kinds of typical events closely related to history and nationality, which play an important role in cultural heritage through generations. However, automatically recognizing cultural events still remains a great challenge since it depends on understanding of complex image contents such as people, objects, and scene context. Therefore, it is intuitive to associate this task with other high-level vision problems, e.g., object detection, recognition, and scene understanding. In this paper, we address this problem by combining both ideas of object / scene contents mining and strong image representation via CNN into a whole framework. Specifically, for object / scene contents mining, we employ selective search to extract a batch of bottom-up region proposals, which are served as key object / scene candidates in each event image; while for representation via CNN, we investigate two state-of-the-art deep architectures, VGGNet and GoogLeNet, and adapt them to our task by performing domain-specific (i.e., event) fine-tuning on both global image and hierarchical region proposals. These two models can complementarily exploit feature hierarchies spatially, which simultaneously capture the global context and local evidences within the image. In our final submission for ChaLearn LAP Challenge ICCV 2015, nine kinds of features extracted from five different deep models were exploited and followed with two kinds of classifiers for decision level fusion. Our method achieves the best performance of $mAP = 0.854$ among all the participants in the track of cultural event recognition.

1. Introduction

An event generally can be defined as a semantically meaningful human activity, taking place within a selected environment and containing a number of necessary objects [12]. As a special case, cultural events are kinds of typical events closely associated with the history and nationality, e.g., *La Tomatina* is a festival that is held in Spain, in which participants throw tomatoes and get involved in this tomato fight purely for entertainment purposes, the *Albuquerque* International *Balloon Fiesta* is a yearly festival of hot air balloons that takes place in New Mexico, USA during early October, and the *Tour de France* is an annual multiple stage bicycle race primarily held in France and occasionally making passes through nearby countries. These cultural events or festivals as cultural heritage are considered to be of great significance for human development and progress. Image as a classic visual media has played an important role to promote the spread of culture with their property that they are easy to store / access / comprehend, especially in the era of the Internet, when more and more photos are continuously uploaded to the Internet via user-generated content websites like *Flickr*, *Facebook*, and *Instagram*. However, the explosive growth of resources makes it almost impossible to manual annotate or tag. Thus, it is necessary to investigate how to understand the cultural events automatically.

In this paper, we formalize this problem as image-based cultural event recognition, i.e., given a cultural event image, assigning it to the class to which it belongs. This is tremendously challenging because images from the same cultural event might have quite different appearances, i.e., large within-class scatter (see Figure 3), moreover, cultural events are complex phenomena which involve interactions among scenes and objects, and therefore analysis of cultural event requires techniques which can go beyond recognizing individual entities and carry out joint reasoning based on evidences of multiple aspects [19]. Based on such requirement, lots of research attempted to solve the event recognition problem by integrating scene and object categorizations [12, 19]. More specifically, [12] utilized a generative graphical model along with a set of integrative and

*The first two authors contributed equally to this work, and were ordered alphabetically.

978-0-7695-5720-5/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCVW.2015.44

hierarchical labels of an image to perform the *what (event), where (scene) and who (object)* recognition of an entire event, and [19] formulated a multi-layer framework to tackle the problem of event recognition, which took into account both visual appearance and the interactions among humans and objects, and combined them via semantic fusion. On the other hand, in recent years, deep convolutional neural network has led to significant progress in several classic vision tasks including object detection [6], object recognition [8], and scene recognition [20], which have been treated as the core technologies of event recognition. Meanwhile, deep convolutional neural network never has been in development from the initial LeNet [11] to AlexNet [8]. More recently, the success of VGGNet (16 or 19 layers) [15] and GoogLeNet (22 layers) [16] implies that deeper structure exhibits a more powerful representation.

In light of such progresses of two streams, we propose to combine both ideas of object / scene contents mining and strong visual representation via CNN into a whole framework. Specifically, for object / scene contents mining, we employ selective search to extract a batch of bottom-up region proposals, which serve as key objects / scene candidates in each event image; while for representation via a CNN, we investigate two state-of-the-art deep architectures, VGGNet [15] and GoogLeNet [16], and adapt them to our task by performing domain-specific (i.e., event) fine-tuning on both global image and hierarchical region proposals. These two models can complementarily exploit feature hierarchies spatially, which simultaneously capture the global context and local evidences within the image. By performing decision level fusion of both models, we can obtain significant improvement compared to the scheme only based on original images. In our final submission for the ChaLearn LAP Challenge ICCV 2015 [4], nine kinds of features extracted from five different deep models were exploited and followed with two kinds of classifiers, i.e., Logistic Regression [5] and Linear Discriminant Analysis [2]. Our method achieves the best performance of $mAP = 0.854$ among all the participants in the track of cultural event recognition. A schema of our proposed method is shown in Figure 1.

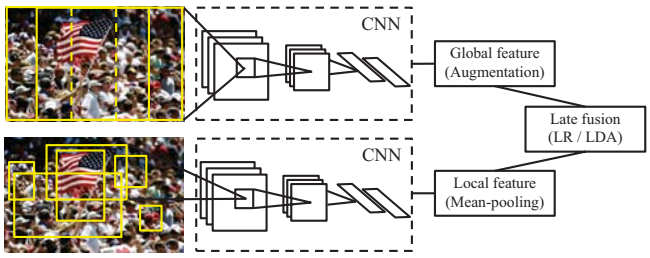

Figure 1. Schema of our proposed method.

The rest of this paper is structured as follows. Section 2 reviews several most related work for our problem and method. Section 3 introduces the whole pipeline including the feature extraction model and classification schemes. In Section 4, implementation details of data preparation and networks structures are provided for the purpose of reproducibility. In Section 5, we report our experimental results with analysis of computational cost. Finally, we conclude our work and discuss possible future efforts in Section 6.

2. Related work

Cultural event recognition is a brand new task raised in ChaLearn LAP Challenge which received several contributions [18, 14, 13, 9] in the CVPR 2015 workshop [1]. Specifically, [18] proposed five-stream object-scene convolutional neural networks to extract important visual cues of both object and scene for event understanding. [14] combined the visual features extracted from convolutional neural networks with metadata (time stamps) of the photos in the hierarchical fusion scheme to generate the final prediction. [13] extracted visual features from a number of subregions within the image, and performed image recognition by combining all the classification results of each subregions probabilistically. [9] directly treat event recognition as a general image categorization problem, and investigated several state-of-the-art methods in this field, e.g., spatial pyramid matching [10], regularized max pooling [7], then least-square SVM was employed on different features, e.g., SIFT, color histogram, and CNN features [8], for final classification. All these efforts mentioned above provided reasonable and effective solutions for the event recognition task, and achieved promising performance ranging from 73% to 85% on the dataset with 50 event categories and 5,875 / 2,332 / 3,569 images for training / validation / testing respectively [1].

Selective search and Region-CNN. Selective search [17] addressed the problem of generating possible object locations for the usage of object recognition. Instead of the sliding window techniques using coarse search grid and fixed aspect ratios, selective search considered the fact that images are intrinsically hierarchical, thus combining bottom-up segmentation and data-driven grouping to generate a compact set of multi-scale subregions for object proposals. Inspired by this, Region-CNN [6] took selective search as a pre-processing module to generate hierarchical region proposals, and then applied high-capacity convolutional neural networks to each region for a much stronger representation, which achieved significant performance boost as well as efficiency for scalable detection task.

3. Feature hierarchies with CNN

The key to event recognition is understanding the complex image contents such as people, objects, and scene con-

text. These contents are intrinsically hierarchical in both aspects of spatial location and semantics, which motivates us to perform multi-scale image partitioning for hierarchical content discovery and representation. Inspired by the recent progresses in object and scene recognition [17, 6, 8, 20], we propose to combine both ideas of hierarchical contents mining and visual representation via CNN into a single framework. In the first step of hierarchical contents mining, we also employ selective search to extract a batch of bottom-up region proposals as in [6], which are served as key objects / scene candidates in certain event image. For each region, we investigate two deep architectures, i.e., VGGNet [15] and GoogLeNet [16] for feature extraction. According to the observation in [6] that when labeled training data is scarce, supervised pre-training for an auxiliary task then followed by domain-specific fine-tuning yields significant improvement, for both GoogLeNet and VGGNet, we perform pre-training on large object dataset, i.e., ImageNet [3] and then fine-tune the model on the training set of cultural event images provided by the challenge, using both schemes of global image and hierarchical region proposals. These two models can complementarily exploit feature hierarchies spatially, which simultaneously capture the global context and local evidences within the image. In Figure 2, we highlight some spatial regions with larger prediction scores (deep networks outputs) in the image, which can be considered as distinctive evidences for the certain event.

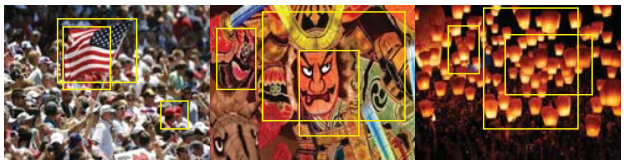

Figure 2. Examples of spatial regions with larger prediction scores (deep networks outputs) in images.

4. Implementation details

4.1. Data preparation

In our framework, we fine-tune our CNN model and extract features based on global images and region proposals respectively. In global scheme, for a single image, if width $>$ height, we resize the image to keep its *height* as 256 pixels and resample it as left, middle, and right parts with the size of 256×256; otherwise, if width $<$ height, we resize the image to keep its *width* as 256 pixels, and resample it as top, middle, and bottom parts, also with the size of 256×256. In training stage, the three parts of each image are all assigned by its image label and involved in the fine-tuning process. While for feature extraction, we average the feature vectors of these three parts to obtain a whole image representation.

For region proposals, we generate a batch of (about 125 per image in our experiments) hierarchical bounding boxes using selective search by filtering out the results whose a) width and height is less than 20% of the original image; b) width/height ratio is greater than 2.0 or less than 0.5. All the sub-regions are then resized to 256×256 directly. As the same as in global scheme, we feed all the region proposals with its image label into the deep networks for fine-tuning, and finally combine the feature vectors of all these regions by mean-pooling to obtain the final representation for further classification.

4.2. Network structure and parameters

In our final submission, two architectures, VGGNet and GoogLeNet are employed for feature extraction. We performs pre-training on ImageNet database and fine-tuning the model in two schemes of global image and region proposals. All the details of the network structures and parameters are summarized as follows:

VGGNet. VGGNet are 16 or 19-layer deep networks with 1 softmax loss layer. For global image, we fine-tune the networks using $42,996$ $(\#train \times 3)$ images with learning rate of 0.001, momentum of 0.9, weight-decay of 0.005, and dropout ratio of 0.5 (for fully-connected layer). The whole process is performed by 30K iterations with mini-batch size of 32; For region proposals, we fine-tune the networks using $1,794,988$ subimages with almost the same parameters, except for the number of iterations of 120K. In both schemes, the output values of the 4096-dimension softmax loss layer are served as the image representation.

GoogLeNet. GoogLeNet is 22-layer deep networks with 3 softmax loss layer. For global image, we fine-tune the networks using $42,996$ images with learning rate of 0.01, momentum of 0.9, weight-decay of 0.005, and dropout ratio of 0.5 (for fully-connected layer). The whole process is performed by 40K iterations with mini-batch size of 128; For region proposals, we fine-tune the networks using $1,794,988$ subimages only with the change of iterations of 100K. The output values of the 1024-dimension softmax loss layer are served as the image representation.

5. Experiments

5.1. Dataset and protocol

In the experiments, we evaluate our method on the ChaLearn Cultural Event Recognition dataset [4]. This dataset contains 28,705 images corresponding to 100 different cultural event categories (99 events and 1 non-class) from all around the globe. The data has been split into three subsets: 14,332 for training, 5,704 for validation, and 8,669 for testing. The distribution of images by category are approximately equal. In all the image categories, human poses, garments, special objects, and scene contex-

Figure 3. Selected samples from five cultural events (*July* 4^{th}, *Tour de France*, *Ballon Fiesta*, *Annual Buffalo Roundup*, and *La Tomatina*) and one non-class (the bottom row). More specifically, *July* 4^{th} is a federal holiday commemorating the adoption of the Declaration of Independence on July 4^{th} in the year of 1776; *Tour de France* is an annual multiple stage bicycle race primarily held in France, while also occasionally making passes through nearby countries; *Ballon Fiesta* is a yearly festival of hot air balloons that takes place in Albuquerque, New Mexico, USA during early October; *Annual Buffalo Roundup* is held in Custer State Park, South Dakota, USA in September, and the bison there are rounded up, with several hundred sold at auction so that the remaining number of animals will be compatible with the rangeland forage; and *La Tomatina* is a festival held in Spain, in which participants throw tomatoes and get involved in this tomato fight purely for entertainment purposes.

t constitute possible cues for characterizing certain events, while preserving the inherent inter / intra class variations. Figure 3 shows some example images for several cultural events, such as *July* 4^{th}, *Tour de France*, *Ballon Fiesta*, *Annual Buffalo Roundup*, and *La Tomatina*.

For evaluation, a precision-recall curve is generated for each category according to the real-valued prediction scores of each image. The principal quantitative measurement is Average Precision (AP), which refers to the area under the precision-recall curve. After the AP's computation of each category, we average all APs to obtain mAP for the final performance. (The evaluation code is provided by ChaLearn LAP Challenge [4].)

5.2. Experimental results

We employ two kinds of linear classifiers, Logistic Regression (LR) and Linear Discriminant Analysis (LDA), on image features extracted from different deep models and fusing their decision scores for final results. For LR, we use the Liblinear package [5] with the parameter "$-s$ 0 $-c$ 1". For LDA, we first conduct PCA for dimension reduction. Specifically, we preserve 3,000 dimensions for VGGNet features and 1,000 dimensions for GoogLeNet features. The final LDA dimension is set as 99, which is one less than the number of categories.

We illustrate all our results based on different models with different classifiers as follows. In Table 1 and Table 2, the comparisons among "(Model) ImageNet", "(Model) ImageRegion", "(Model) ImageNetEvents", and "(Model) ImageNetEventsRegion" demonstrate the effectiveness of both operations of region proposals and domain-specific fine-tuning. Table 3 shows the fusion results based on nine different CNN features and two classifiers, and we achieve the $mAP = 0.850$ on validation set. Finally, the challenge results on testing set provided by organizers are shown in Table 4.

Table 1. Performance on validation set based on VGGNet.

Models	LR	LDA
VGG16ImageNet	0.639	0.648
VGG16ImageNetRegion	0.707	0.697
VGG16ImageNet (*finetune*)	0.735	0.741
VGG16ImageNetRegion (*finetune*)	0.786	0.793
VGG19ImageNet	0.626	0.640
VGG19ImageNetRegion	0.709	0.695
VGG19ImageNet (*finetune*)	0.728	0.734
VGG19ImageNetRegion (*finetune*)	0.782	0.790

**1. Model: [Networks] [Dataset] ... [Region (optional)]*
**2. [Region]: Region proposals generated by selective search.*

5.3. Computation time

In this section , we report the computation time of each module in our framework. In feature extraction step, we only consider the fine-tuning cost based on the pre-trained networks model. For

Table 3. Performance on validation set based on multiple models fusion.

Models	LR	LDA	Fusion LR+LDA
GoogleImageNetRegion (*finetune*)	0.805	0.804	0.820
+ GoogleImageNetRegion (*finetune*) (*loss1* + *loss2*)	0.813	0.804	0.824
+ (VGG16 + VGG19) ImageNetRegion (*finetune*)	0.830	0.826	0.839
+ (VGG16 + VGG19) ImageNet (*finetune*)	0.841	0.831	0.846
+ (VGG16 + VGG19) ImageNetRegion	0.845	0.829	0.850

Table 2. Performance on validation set based on GoogLeNet.

Models	LR	LDA
GooglePlaces	0.505	0.416
GooglePlaces (*finetune*)	0.689	0.708
GoogleImageNet	0.551	0.537
GoogleImageNet (*finetune*)	0.723	0.739
GoogleImageNetRegion (*finetune*)	0.805	0.804
GoogleImageNetRegion (*finetune*) (*loss1*)	0.753	0.758
GoogleImageNetRegion (*finetune*) (*loss2*)	0.788	0.793

Table 4. Performance on testing set from all participants.

Position	Team	Development	Test
1	VIPL-ICT-CAS	0.783	**0.854**
2	FV	0.770	0.851
3	MMLAB	0.717	0.847
4	NU&C	0.387	0.824
5	CVL_ETHZ	0.662	0.798
6	SSTK	0.740	0.770
7	MIPAL_SNU	0.801	0.763
8	ESB	0.729	0.758
9	UPC-STP	0.503	0.588

global images, we fine-tune the networks using $42,996$ ($\#train\times 3$) images with 30K iterations on VGGNet and 40K iterations on GoogLeNet respectively, each of which takes about 1 day with Tesla K40 GPU. For region proposals, we fine-tune the networks using $1,794,988$ subimages with 120K iterations on VGGNet and 100K iterations on GoogLeNet, which takes about 2 days and 3 days respsctively with Tesla K40 GPU. For classification, we summarize the train / test expended time of both validation stage and final testing stage in Table 5 and Table 6, regarding to different deep models and different classifiers. (Note that all data are obtained using one PC with 2.20GHz and 4G RAM.)

6. Conclusions

In this paper, we present our method for cultural event recognition by exploiting visual feature hierarchies with deep convolutional neural networks. Technically, for each cultual event image, selective search is first conducted for region proposals extraction, then both global image and subimage regions are served as input

Table 5. Computation time (train / test) in validation stage.

Models	LR		LDA	
	train	test	train	test
GoogLeNet	214.39s	0.82s	7.80s	10.11s
VGGNet	379.25s	1.12s	146.53s	10.37s

Table 6. Computation time (train / test) in test stage.

Models	LR		LDA	
	train	test	train	test
GoogLeNet	424.02s	1.17s	9.08s	32.48s
VGGNet	587.86s	1.65s	146.53s	37.92s

to fine-tune two deep convolutional neural networks for hierarchical feature learning. At the back-end, we simply utilize two classic discriminant learning methods for classification and perform decision-level fusion for final predictions. In the future work, we'll try to consider more complex visual cues, like human poses, garment, or human-object interactions, and figure out their intrinsic relationship semantically for in-depth understanding of the scenes and events.

Acknowledgements

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61222211, 61379083, and National Science Foundation under Grant No. IIS-1251187.

References

[1] X. Baró, J. Gonzàlez, J. Fabian, M. A. Bautista, M. Oliu, H. J. Escalante, I. Guyon, and S. Escalera. Chalearn looking at people 2015 challenges: action spotting and cultural event recognition. In *CVPRW*, 2015.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE T PAMI*, 19(7):711–720, 1997.

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

[4] S. Escalera, J. Fabian, P. Pardo, X. Baró, G. Jordi, H. J. Escalante, and G. Isabelle. Chalearn 2015 apparent age and

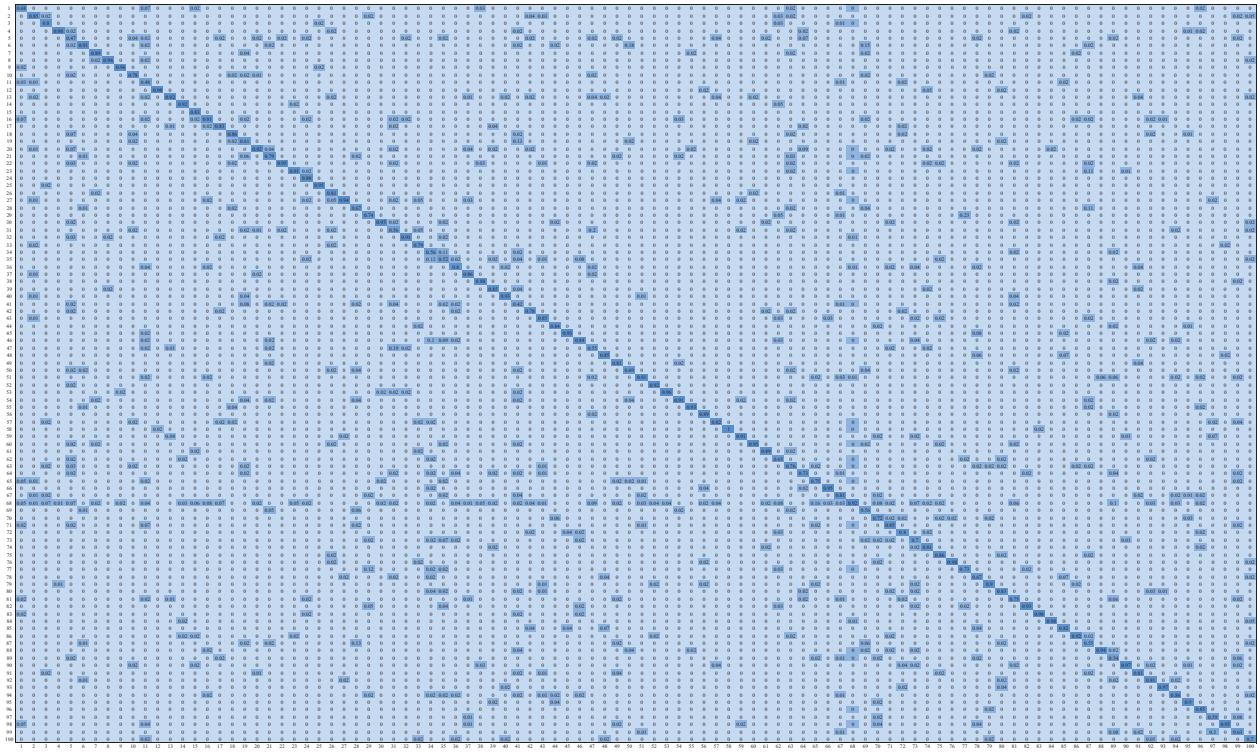

Figure 4. Confusion matrix of the 100 cultural events (100 rows by 100 columns). Please kindly note that some events could be easily confused with each other, e.g., *Eid al-Adha* (34^{th} row) v.s. *Eid al-Fitr Iraq* (35^{th} column), *Carnaval de Oruro* (19^{th} row) v.s. *Fiesta de la Candelaria* (41^{th} column), and *Asakusa Samba Carnival* (6^{th} row) v.s. *Notting hill carnival* (69^{th} column). Moreover, there are also several cultural events having quite satisfactory accuracy, e.g., *Pingxi Lantern Festival* (76^{th}), *Sandfest* (84^{th}), and *Aomori Nebuta* (4^{th}). Such high accuracies are mainly attributed to the small intra-class variance. Besides, it is obvious that the *non-class* (68^{th}) has high confusion score with most of other events.

cultural event recognition: datasets and results. In *ICCVW*, 2015.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[7] M. Hoai. Regularized max pooling for image categorization. In *BMVC*, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[9] H. Kwon, M. Hoai, K. Yun, and D. Samaras. Recognizing cultural events in images: a study of image categorization models. *CVPRW*, 2015.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

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

[12] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[13] S. Park and N. Kwak. Cultural event recognition by subregion classification with convolutional neural network. *CVPRW*, 2015.

[14] A. Salvador, M. Zeppelzauer, D. Manchon, A. Calafell, and X. Giro-Nieto. Cultural event recognition with visual convnets and temporal models. *CVPRW*, 2015.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv*, 2014.

[17] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

[18] L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. *CVPRW*, 2015.

[19] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In *CVPR*, 2015.

[20] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.