指导教师：杨涛　　提交时间：2016.3.17

CVPR2015 Paper Translation

No: 01
姓名：高雅濛
学号：2013302515
班号：10011303

利用特征层次结构与卷积神经网络进行文化活动识别

Mengyi Liu*1, Xin Liu*1, Yan Li1, Xilin Chen1, Alexander G. Hauptmann2,Shiguang Shan1

1Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),

Institute of Computing Technology, CAS, Beijing, 100190, China

2School 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

摘要

文化活动是一种典型的与历史和国籍紧密相关的活动，它在一代又一代文化遗产的传承中起着十分重要的作用。然而，自动识别文化活动仍然是一个巨大的挑战，因为它取决于对复杂的图像内容的理解，如人、对象、以及现场环境。因此，可以直观的将这个工作与其他高层次的视觉问题联系起来，例如，目标检测、识别和场景理解。在本文中，我们利用将对象/场景内容挖掘的思想和强大的图像结合起来通过 CNN 表示成一个整体框架的方法来解决这个问题。特别地，对于对象/场景内容的挖掘，我们采用选择性搜索来提取一批自下而上的地区提案，来作为在每个活动图像中的关键对象/场景的候选；而通过 CNN 的表示，我们研究两个最先进的深层架构，VGGNet 和 GoogLeNet，并且使他们通过执行特定领域的任务（也就是活动）和整合全局的图像以及分层的地区提案来适应我们的工作。这两个模型可以利用特征层次结构空间进行互补，同时也可以捕获全局上下文和本地图像中的证据。在我们最终提交于 ChaLearn LAP ICCV 2015 挑战赛的文章中，我们从五个不同深度的模型中提取了九种特征，并利用和遵循两种分类器对其进行决策级的融合。我们的方法在所有对文化活动进行跟踪识别的参与者中达到了最好的效果，即 mAP = 0.854。

1. 介绍

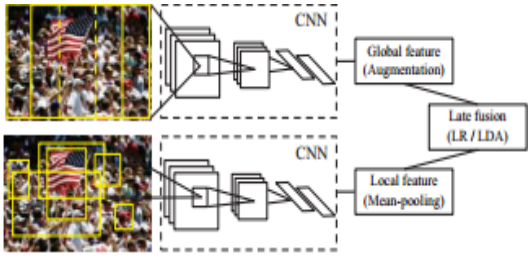
一个事件通常可以被定义为一个语义上有意义的人类活动，它发生在一个选定的环境中并且包含大量的必要的对象【12】。作为一种特殊的情况，文化活动是一种典型的与历史和国籍紧密相关的事件，例如，狂欢节是西班牙的一个节日，庆祝节日的参与者们互扔番茄，他们参与这场番茄大战纯粹是为了娱乐；阿尔伯克基国际气球节是一个一年一度的热气球节日，新墨西哥和美国在每年十月初举行该节日；环法自行车赛是一个年度多级自行车竞赛主要在法国举行，偶尔会穿过附近的国家。这些文化活动或节日文化遗产被认为对人类的发展和进

步具有重要的意义。图像作为一个经典的视觉多媒体对促进文化以及他们的所有权的传播发挥了重要作用，正因如此，图像很容易被存储/访问/理解，尤其是在互联网时代，当越来越多的照片通过许多用户生成的网站不断地被上传到互联网，如 Flickr,Facebook 和 Instagram。然而，资源的爆炸性增长使得它几乎不可能进行手工注释或标记。因此，有必要研究如何自动理解文化活动。

在本文中，我们基于对图像的文化活动识别来研究这个问题，也就是说，给一个文化活动的图像，将它分配给它所属的类。这是非常具有挑战性的，因为来自同一个文化活动的图像可能会有完全不同的表象，即大的样本类内离散度（见图 3），此外，文化活动是涉及场景和对象之间的交互的复杂的现象，因此分析文化活动需要能够超越单个实体识别和事项基于多个方面证据的联合推理的技术【19】。基于这样的需求，许多研究试图通过集成场景和对象分类的方法来解决事件识别问题【12,19】。更具体地说，【12】利用图形化生成模型和一个图像的一组综合和分层标签来实现对整个事件的活动，场景，对象的识别，【19】并制定一个多层框架来解决活动识别的问题，既考虑了视觉外观以及人类和对象之间的交互，又将他们通过语义进行融合。另一方面，近年来，深卷积神经网络使一些经典的视觉任务取得了重大的进展，包括目标检测【6】、目标识别【8】、场景识别【20】，并且场景识别已经被视为活动识别的核心技术。与此同时，深卷积神经网络从最初的 LeNet【11】到 AlexNet【8】一直都在发展中。最近，VGGNet（16 层到 19 层）【15】和 GoogLeNet（22 层）【16】的成功意味着深层次结构的一个更强大的展示。

鉴于这样的两个流的进展，我们建议将对象/场景内容挖掘的思想和强烈的视觉结合起来通过 CNN 表示成一个整体框架。特别地，对于对象/场景内容的挖掘，我们采用选择性搜索来提取一批自下而上的地区提案，来作为在每个活动图像中的关键对象/场景的候选；而通过 CNN 的表示，我们研究两个最先进的深层架构，VGGNet【15】和 GoogLeNet【16】，并且使他们通过执行特定领域的任务（也就是活动）和整合全局的图像以及分层的地区提案来适应我们的工作。这两个模型可以利用特征层次结构空间进行互补，同时也可以捕获全局上下文和本地图像中的证据。通过执行融合了两种模型的决策级，我们可以获得显著的改善相比于仅基于原始图像的方法。在我们最终提交于 ChaLearn LAP ICCV 2015【4】挑战赛的文章中，我们从五个不同深度的模型中提取了九种特征，并利用和遵循两种分类器，即，逻辑回归【5】和线性判别分析【2】。我们的方法在所有对文化活动进行跟踪识别的参与者中达到了最好的效

果，即 mAP = 0.854。一个被我们推荐的方法的模式如图 1 所示。



图一 推荐的方法的模式

本文其余部分的结构如下。第二部分回顾了大多数与我们的问题相关的工作和方法。第三部分介绍了整个流水线包括特征提取模型和分类方案。在第四部分，数据准备的细节和网络结构的实现提供了再现性。在第五部分中，我们报告了实验结果并分析了计算成本。最后，在第六部分我们总结了我们的工作并讨论了未来可能要做出的努力。

2. 相关工作

文化活动识别在 ChaLearn LAP 挑战赛中是一项全新的任务，它在 2015 年 ICCV 研讨会【1】中收获了几个贡献【18, 14, 13, 9】。具体来说，【18】提出了单条流的对象-场景卷积神经网络来提取出对活动理解有重要作用的对象和场景的视觉线索。【14】结合从卷积神经网络提取的视觉特征和在分层融合方案中的照片的元数据（时间戳）来生成最终的预测。【13】从图像许多条件中提取视觉特征，并结合每个条件概率问题的分类结果进行图像识别。【9】直接把活动识别问题看作一般的图像分类问题，并研究了一些在此领域先进的方法，例如，空间金字塔匹配【10】、正规化 max 池【7】、在不同的特性上采用最小二乘支持向量机，例如，对最终分类进行筛选、得到颜色直方图以及 CNN 特性【8】。所有上面所提到的努力都是为活动识别任务提供合理有效的解决方案，实现了承诺的性能包括从 73%到 85%不等的数据集和 50 种活动类别，以及对 5,875/2,332/3,569 图片的训练/验证/测试【1】。

选择搜索和基于区域的卷积神经网络。选择搜索【17】使用对象识别来解决可能产生的对象的位置的问题。代替滑动窗口技术而使用粗搜索网格和固定的纵横比，选择搜索考虑的实际上是图像的内在层次，因此结合自下而上的分割和数据驱动的分组为对象生成一组紧凑的多尺度条件的提案。受这个的启发，基于区域的卷积神经网络【6】用选择搜索作为一个预处理模块生成区域分层方案，然后应用高容量卷积神经网络使每个区域都获得比之前强得多的代表权，这不仅实现了性能的显著提升，还实现了效率可伸缩的检测任务。

3. 功能层次与卷积神经网络

活动识别的关键是理解复杂的图像内容如人，对象和场景上下文。这些内容是空间位置和语义两方面的内在层次，它激励我们为了层次内容的发现和表示进行多尺度图像分割。灵

感来自于最近发展的对象和场景的识别[17,6,8,20]，我们建议将层次内容挖掘的思想和可视化表示结合起来通过 CNN 生成一个单一的框架。在层次内容挖掘的第一步，我们仍采用选择搜索提取一些自下而上的区域方案作为在某些活动图像中关键的对象/场景的候选[6]。对于每个区域，我们研究两个深层架构，即对 VGGNet[15]和 GoogLeNet[16]进行特征提取。据观察[6]，[当标签的训练数据不足时，监督训练的一个辅助任务，然后进行特定区域的微调可使收益率显著改善，对于 GoogLeNet 和 VGGNet，我们对大的对象数据集进行训练，即 ImageNet[3]和根据提供的挑战利用全局图像和分层区域的方案调整正在训练的文化活动的图像集。这两个模型可以利用特征层次结构空间进行互补，同时也可以捕获全局上下文和本地图像中的证据。在图二中，我们将重点介绍一些存在空间限制的区域和更大空间的预测分数（深层网络输出）的图像，它可以被视为是确定一种活动的独特的依据。

图二　图像中存在空间限制的区域和更大空间的预测分数（深层网络输出）的例子

4. 实现细节

4.1. 数据准备

在我们的框架中，我们分别基于全局图像和区域方案进行 CNN 模型的调整和特征的提取。在全局计划中，对于单个的图像，如果宽度>高度，我们调整图像使它的高度保持在 256 个像素，并以 256*256 的大小对图像的左边、中间以及右边部分重新取样；否则，如果宽度<高度，我们调整图像使它的宽度保持在 256 个像素，并以 256*256 的大小对图像的上部、中间部分以及下部重新取样。在训练阶段，每个图像的三个部分都是由其图像标签和参与微调的进程来完成的，而对于特征提取，我们平均这三个部分的特征向量来获得一个整体的图像表示。对于区域方案，我们生成一批层次边界框（在实验中为每张图像 125 个）使用选择搜索通过过滤的方法过滤掉 a)宽度和高度小于原始图像 20%的结果；b)宽度/高度比率大于 2.0 或小于 0.5。所有的次区域都被调整为 256*256 像素。与全局方案一样，我们将所有的区域方案及它的图像标签注入到深层网络进行微调，最后通过对领域内特征点只求平均的方法结合所有这些区域的特征向量来为进一步分类获得最终表示。

4.2. 网络结构和参数

在我们最终提交中，VGGNet 和 GoogLeNet 两个架构用于特征提取。我们对 ImageNet 数据库进行训练并全局

图像和区域图像两个方案中的模型进行微调。网络结构和参数的所有细节总结如下：

VGGNet。VGGNet 是有一层 softmax 损失层的 16 或 19 层的深度网络层。对于全局图像，我们使用学习速率为 0.001、冲力为 0.9、重量衰减为 0.005、释放比率为 0.5（对于全层）的 42,996（#train*3）张图像对网络进行调整。整个过程是由大小为 32 的小批量的 30k 次迭代完成的；对于区域方案，我们使用 1,794,998 张有几乎相同参数（除了 120k 次迭代）的子图像来调整网络。在这两种方案中，4096 维的 softmax 损失层的输出值被用于图像的表示。

GoogLeNet。GoogLeNet 是有三层 softmax 损失层的 22 层深度网络层。对于全局图像，我们使用学习速率为 0.01、冲力为 0.9、重量衰减为 0.005、以及释放比率为 0.5（对于全层）的 42,996 张图像对网络进行调整。整个过程是由大小为 128 的小批量的 40k 次迭代完成的；对于区域方案，我们仅仅使用有 100k 次迭代的子图像来调整网络。1024 维的 softmax 损失层的输出值被用于图像表示。

图三　从五种文化活动中选择样本（七月四日，环法自行车赛，气球嘉年华，一年一度的水牛节和番茄狂欢节）和一个其他类型的（最后一行）。更具体地说，七月四日是一个联邦假日为了纪念 1776 年 7 月 4 日《独立宣言》被正式采用；环法自行车赛是一个年度多级自行车竞赛，主要在法国举行，偶尔也穿过附近的国家；阿尔伯克基国际气球节是一个一年一度的热气球节日，新墨西哥和美国在每年十月初举行该节日；一年一度的水牛节在卡斯特州立公园、南达科他、美国九月份举行，水牛被围捕，数百头进行拍卖，这样牧场饲料将足够喂养剩下的动物；番茄狂欢节是一个在西班牙举行的节日，这一天参与者们互扔番茄并且他们参与这场番茄大战纯粹是为了娱乐。

5. 实验

5.1. 数据集和协议

在实验中，我们评估我们的方法基于 ChaLearn 文化活动识别的数据集[4]，这个数据集包含 28,705 张图像对应于来自世界各地的 100 个不同的文化活动类别（99 个活动和 1 个其他类型的）。将数据分成三个子集：14,332 个用于训练，5，704 用于验证，8,669 用于测试。图像分布的类别大致相等。在所有的图像类别中，人的姿势、服装、特殊对象和场景上下文为描述某些活动构成可能的线索，同时保留固有的内部类变化。图三中展示了一些文化活动的图像，如 7 月 4 日、环法自行车赛、一年一度的水牛节、气球嘉年华以及番茄狂欢节。

对于评估，precision-recall 曲线根据每个图像的实值预测分数来生成每个类别。定量测定的方法为求取平均精度（AP），它是指在 precision-recall 曲线下的面积，在用 AP 对每个类别进行计算之后，我们对所有的 AP 取平均值得到最终性能的 mAP（评估代码是由 ChaLearn LAP 挑战赛[4]提供的）。

5.2. 实验结果

我们采用两种线性分类器，逻辑回归分类器（LR）和线性判别分析分类器（LDA）来从不同深度模型提取图像特征并融合他们的决策分数作为最终结果。对于逻辑回归分类器，我们使用参数为“-s 0 -c 1”的 Liblinear 包[5]。对于线性判别分析分类器，我们首先实施 PCA 进行降维。具体来说，我们保留 3，000 维作为 VGGNet 架构的特征，保留 1,000 维作为 GoogLeNet 架构的特征。最终的 LDA 维度被设置为 99，它的维度小于大多数类别的维度。

下面我们根据不同的模型和分类对所有结果进行说明。在表 1 和表 2 中，对“ImageNet”、“ImageRegion”、“ImageNetEvents”和“ImageNetEventsRegion”几种模型之间进行了比较，展示了区域方案操作和特定领域微调的有效性。表 3 展示了基于九种不同的 CNN 特征和两种分类器的融合结果，并且我们在验证集中实现了 mAP=0.850。最后，主办方提供的测试挑战结果如表 4 所示。

表三　基于多模型融合验证集的性能

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

表一 基于 VGGNet 验证集的性能

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.*

表二 基于 GoogLeNet 验证集的性能

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

表四 对所有参与者测试得到的性能

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

表五 验证阶段（训练/测试）的计算时间

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

表六 最终测试阶段（训练/测试）的计算时间

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

5.3. 计算时间

在这个部分中，我们报告我们的框架中每个模块的计算时间。在特征提取步骤中，我们只考虑基于预训练网络模型的调整成本。对于全局图像，我们分别使用在 VGGNet 架构上进行了 30k 次迭代的 42,996（#train*3）张图像和在 GoogLeNet 架构上进行 40k 次迭代来调整网络，每个大约需要 Tesla 40K GPU 计算一天。对于区域方案，我们分别使用在 VGGNet 架构上进行了 120k 次迭代的 1,794,998 张子图像和在 GoogLeNet 架构上进行 100k 次迭代来调整网络，这分别需要 Tesla 40K GPU 计算两天和三天。对于分类，在表 5 和表 6 中我们总结了验证阶段和最终测试阶段训练/测试花费的时间，对于不同深度的模型和不同的分类器（注意：所有的数据都是由 2.20GHz 和 4G RAM 的计算机得到的）。

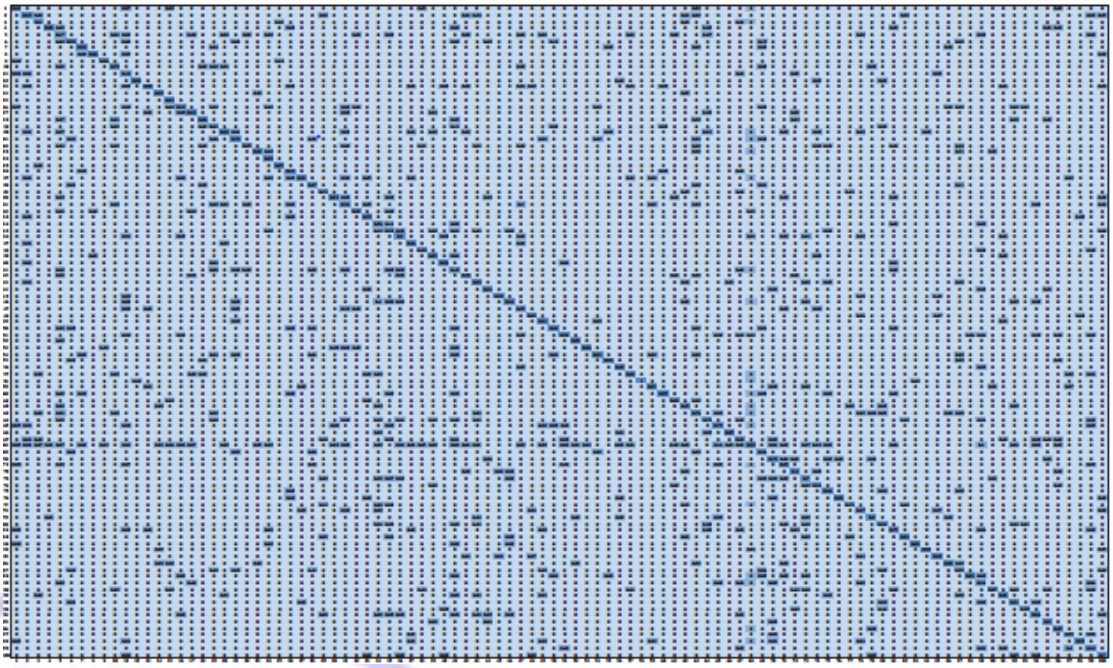
6. 结论

在本文中，我们通过利用视觉特征层次结构与深卷积神经网络层次结构来提出文化活动识别的方法。从技术上来讲，对每个文化活动图像，首先使用选择搜索进行区域方案提取，然后全局图像和区域子图像都是作为输入来调整两个深卷积神经网络，从而进行特征层次结构的学习。最后，我

们仅仅利用两个经典判别学习方法进行分类和为了最终预测执行的决策级融合。在未来的工作中，我们将会尝试去考虑更复杂的视觉线索，像人的姿势、服装或是人与物体之间的互动，并找出其内在关系来对场景和活动有深入的理解。

鸣谢

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.



图四　容易混淆的 100 种文化活动的矩阵（100 行、100 列）。请注意：一些活动很容易和其他活动混淆，例如，宰牲节（第 34 行）和伊拉克的开斋节（第 35 列）、德欧鲁罗的狂欢节（第 19 行）和坎德拉利亚的嘉年华（第 14 列）、桑巴狂欢节（第 6 行）和诺丁山狂欢节（第 69 列）。此外，还有一些文化活动有相当令人满意的精度值，例如，元宵节（第 76 位）、Sandfest（第 84 位）、Aomori Nebuta（第 4 位）。得到如此高的精度主要归功于方差小的内部类。此外，很明显可以看出其他类型（第 68 位）与大多数其他的活动有较高的混淆值。

参考文献

[1] X. Baro, J. Gonz ´alez, 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. Baro, G. Jordi, H. J. Es- calante, and G. Isabelle. Chalearn 2015 apparent age andcultural 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. Gradientbased 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 scenerecognition using places database. In *NIPS*, 2014.