❖特约专栏❖

基于多层次特征差异图的视觉场景识别

张国山,张培崇,王欣博

(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:场景外观剧烈变化引起的感知偏差和感知变异给视觉场景识别带来了很大的挑战。现有的利用卷积神经网络(CNN)的视觉场景识别方法大多数直接采用 CNN 特征的距离并设置阈值来衡量两幅图像之间的相似性,当场景外观剧烈变化时效果较差,为此提出了一种新的基于多层次特征差异图的视觉场景识别方法。首先,一个在场景侧重的数据集上预训练的 CNN 模型被用来对同一场景中感知变异的图像和不同场景中感知偏差的图像进行特征提取。然后,根据 CNN 不同层特征具有的不同特性,融合多层 CNN 特征构建多层次特征差异图来表征两幅图像之间的差异。最后,视觉场景识别被看作二分类问题,利用特征差异图训练一个新的 CNN 分类模型来判断两幅图像是否来自同一场景。实验结果表明,由多层 CNN 特征构建的特征差异图能很好地反映两幅图像之间的差异,文中提出的方法能有效地克服感知偏差和感知变异,在场景外观剧烈变化下取得很好的识别效果。

关键词:视觉场景识别; 特征差异图; 感知偏差; 感知变异; 卷积神经网络

中图分类号:TP391.4 **文献标志码**:A **DOI**:10.3788/IRLA201847.0203004

Visual place recognition based on multi-level feature difference map

Zhang Guoshan, Zhang Peichong, Wang Xinbo

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Perceptual aliasing and perceptual variability caused by drastically appearance changing in the scene bring great challenge to visual place recognition. Many existing visual place recognition methods using CNN directly adopted the distance of the CNN features and set thresholds to measure the similarity between the two images, which had shown a poor performance when drastically appearance changing in the scene. A novel multi-level feature difference map based visual place recognition method was proposed. Firstly, a CNN pretrained on scene-centric dataset was adopted to extract features for perceptually different images of same place and aliased images of different places. Then, according to the different properties of different CNN layers, multi-level feature difference map was constructed on the multi-level CNN features to represent the difference between the two images. Finally, visual place recognition was regarded as a binary classification task. The feature difference maps were used to train a

收稿日期:2017-10-05; **修订日期**:2017-12-11

基金项目:国家自然科学基金(61473202)

作者简介:张国山(1961-),男,教授,博士生导师,博士,主要从事线性与非线性系统控制、智能控制、图像处理等方面的研究。Email:zhanggs@tju.edu.cn

通讯作者:张培崇(1992-),女,硕士生,主要从事深度学习方面的研究。 Email: 13703932741@163.com

new CNN classification model for determining whether the two images are from the same place. Experimental results demonstrated that the feature difference map constructed by multi-level CNN features can well represent the difference between two images, and the proposed method can effectively overcome perceptual aliasing and perceptual variability, and achieve a better recognition performance when drastically appearance changing in the scene.

Key words: visual place recognition; feature difference map; perceptual aliasing; perceptual variability; convolutional neural network

0 引 言

给定场景的一幅图片,人类或者机器人可以判断出这幅图片是否来自之前见到的场景,这就是视觉场景识别要解决的问题。虽然视觉场景识别已经取得了很大的进展,但是由场景外观剧烈变化引起的感知偏差和感知变异依然给场景识别带来了很大的挑战。一个鲁棒的视觉场景识别系统必须能够成功地匹配来自同一场景但视觉上容易混淆为不同场景的两幅图像(感知变异),同时拒绝来自不同场景但视觉上容易混淆为相同场景的两幅图像(感知偏差)[1]。

一个条件不变的特征描述子对视觉场景识别至关重要。传统的图像特征描述子可以分为局部特征和全局特征。局部特征描述子提取图像的感兴趣区域,而全局特征描述子把图像看作一个整体忽略其内容。SIFT (Scale-Invariant Feature Transforms)[2]和SURF(Speeded-Up Robust Features)[3]是两种常见的局部特征描述子,具有光照、尺度等不变性。FAB-MAP(Fast Appearance Based Mapping)[4]采用词袋模型(Bag-of-Words,BoW)[5]构建局部特征描述子,词袋模型可以把局部特征量化到词典中进而采取文本检索的技术。Vocabulary trees[6]利用分层的模型来定义单词,在大场景的识别中加快了搜索速度。Gist[7]是一个全局特征描述子,对整幅图像在不同的方向和频率上采用 Gabor 滤波来提取特征。图像直方图[8]具有旋转不变性,也是一种广泛应用的全局特征。虽然传统特征描述子取得了很好的识别效果,但仍具有很多缺点。局部特征忽略了场景的空间和结构信息,极大地限制了它的性能,而全局特征更易受到相机视角的影响。

近几年,CNN 在图片分类[9]、图片检索[10]、目标检测[11]、目标跟踪[12]、图像校正[13]等领域有了重大的突破。基于 CNN 的视觉场景识别可以分为两类:一类是直接采用预训练的模型;另一类是采用端到端的学习方式。Chen 等人采用在 ImageNet 数据集上训练的 Overfeat 模型提取图像特征,之后利用两个连续滤波降低误检率。Li 等人[14]先对图像进行超像素分割,然后基于深层 CNN 特征得到超像素描述子,最后提出一种自适应的加权图像相似度衡量方法进行场景识别。Hou 等人[15]采用在 Places[16]数据集上训练的模型进行闭环检测,在光照发生变化的场景下显著优于传统的特征。Arandjelovic 等人[17]在 CNN 结构中添加了一个类似 VLAD (Vector of Locally Aggregated Descriptors)[18]的层“NetVLAD”,并且采用弱监督的学习方式实现端到端的训练。这些方法与基于传统特征的方法相比取得了很好的效果,但大多数利用CNN特征的距离并设置阈值来衡量图片的相似性,当光照、视角等引起感知偏差和感知变异时效果较差。

文中提出了一种基于多层次特征差异图的视觉场景识别方法。首先,利用在场景侧重而不是对象侧重的数据集上预训练的 CNN 模型进行特征提取;然后,通过对单层特征构建特征差异图的性能分析,融合不同层的 CNN 特征构建多层次特征差异图;最后,设计一个新的适合训练特征差异图的分类模型得到最终的识别结果。实验结果表明融合不同层特征的特征差异图能很好地表征图像之间的差异,在发生感知偏差和感知变异的数据集上取得很好的识别效果。

1 基于多层次特征差异图的视觉场景识别

文中提出的方法主要包括三个步骤:特征提取、构建特征差异图和训练分类模型进行场景识别。方法的流程图如图 1 所示。

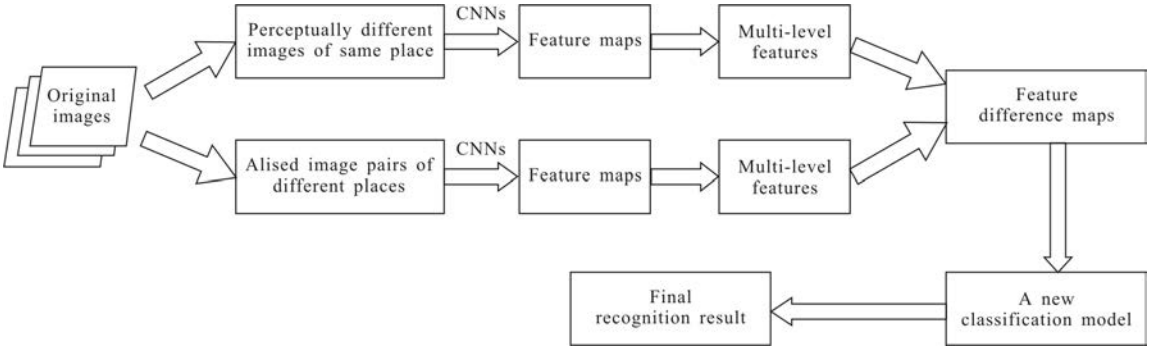


图 1 基于多层次特征差异图的视觉场景识别的流程图

Fig.1 Flow of visual place recognition based on multi-level feature difference map

1.1 特征提取

训练集的类型和网络结构是采用 CNN 进行特征提取的两个重要考虑因素。首先，对象侧重(Object-centric)的数据集，例如 ImageNet，对分类任务具有很好的效果，但并不适合识别任务。场景侧重(Scene-centric) 的数据集，例如，Places 和 Google Street View dataset 更适合识别任务。其次，CNN 不同层的特征具有不同的特性，中等层次的特征包含较多几何信息，对光照等的变化具有较好的鲁棒性，而高层次特征则包含更多的语义信息，能有效地克服视角的变化[19]。因此，选择在 Places 数据集上训练的多层 CNN 模型 vgg16[20]进行特征提取。该网络包含 5 个卷积阶段和 3 个全连接阶段，前两个卷积阶段分别包含 2 个卷积层，2 个 ReLU 层和 1 个池化层，剩下的 3 个卷积阶段分别包含 3 个卷积层，3 个 ReLU 层和 1 个池化层；前两个全连接阶段分别包含 1 个全连接层，1 个 ReLU 层和 1 个 dropout 层，最后 1 个全连接阶段只包含 1 个全连接输出层。

$$F_k(I)=(x_{k1},\cdots,x_{kd}) \tag{1}$$

式中：$F_k(I)$ 表示图片 I 通过 CNN 得到的第 k 层特征。

1.2 单层 CNN 特征构建特征差异图

通过特征提取已经获得了图像对$\{I_1,I_2\}$的第 k 层特征$\{F_k(I_1),F_k(I_2)\}$，为了分析不同层 CNN 特征构建特征差异图的性能，下面利用单层 CNN 特征构建特征误差图。

$\mathrm{diff}_k\{I_1,I_2\}$表示图像对$\{I_1,I_2\}$第 k 层特征的差异，定义如下：

$$\mathrm{diff}_k\{I_1,I_2\}=F_k(I_1)-F_k(I_2)=(y_{k1},\cdots,y_{kd}) \tag{2}$$

式中：d 为特征向量的维数。

之后对 $\mathrm{diff}_k\{I_1,I_2\}$采取标准化操作如下：

$$\mathrm{std-diff}_k\{I_1,I_2\}=255*\frac{|\mathrm{diff}_k\{I_1,I_2\}|}{\max\{y_{k1},\cdots,y_{kd}\}} \tag{3}$$

为了得到合适的特征差异图作为分类模型的训练集，$\mathrm{std-diff}_k\{I_1,I_2\}$需要转化为大小合适的图片，转化后的大小如表 1 所示。

表 1 特征向量的维数和特征差异图的大小

Tab.1 Dimension of the feature vector and the size of the feature difference map

Layer	Dimension	Size
conv3_	802 816	784×1024
pool3	200 704	256×784
conv4_	401 408	512×784
pool4	100 352	256×392
conv5_	100 352	256×392
pool5	25 088	128×196
fc	4 096	64×64

1.3 多层 CNN 特征构建特征差异图

为了融合不同层的特征，需要将不同大小的 feature maps 转化为相同的尺寸。对较大尺寸的 feature maps 采用 max pooling 和卷积，对较小尺寸的 feature maps 采用反向 pooling 和反卷积(deconvolution)[21]。图 2 给出了融合 conv3_3，conv4_3 和 conv5_3 的示例，conv3_3 首先通过 max pooling 得到 28×28×256 的 feature maps，然后利用 3×3×512 的卷积转化为 28×28×512 的 feature maps；conv4_3 保持不变；conv5_3 通过三次反卷积和一次反向 pooling 得到 28×28×512 的

feature maps。

$\widetilde{F}_{conv3\_3}(I)$,$\widetilde{F}_{conv4\_3}(I)$,$\widetilde{F}_{conv5_3}(I)$分别表示 conv3_3,conv4_3,conv5_3 处理之后的特征向量,其大小均为 512×28×28。$\widetilde{F}_{conv33-conv43-conv53}(I)$表示融合该三层得到的特征,计算如下:

$$\widetilde{F}_{conv33-conv43-conv53}(I)=\omega_{conv3\_3}\widetilde{F}_{conv3\_3}(I)+\omega_{conv4\_3}\widetilde{F}_{conv4\_3}(I)+\omega_{conv5\_3}\widetilde{F}_{conv5\_3}(I) \tag{4}$$

式中:ω_{conv3_3},ω_{conv4_3},ω_{conv5_3} 分别表示对应的权重,$\omega_{conv3\_3}+\omega_{conv4\_3}+\omega_{conv5\_3}=1$ 且其值参考对应的单层特征得到的特征差异图的识别性能设置为常量。

$diff_{conv33-conv43-conv53}\{I_1,I_2\}$表示利用融合后特征计算出的图像对$\{I_1,I_2\}$之间的差异,计算如下:

$$diff_{conv33-conv43-conv53}\{I_1,I_2\}=\widetilde{F}_{conv33-conv43-conv53}(I_1)-\widetilde{F}_{conv33-conv43-conv53}(I_1) \tag{5}$$

之后进行标准化,最后把标准化得到的$diff_{conv33-conv43-conv53}\{I_1,I_2\}$转化为 512×784 的图片,即为融合 conv3_3,conv4_3,conv5_3 得到的多层次特征差异图。

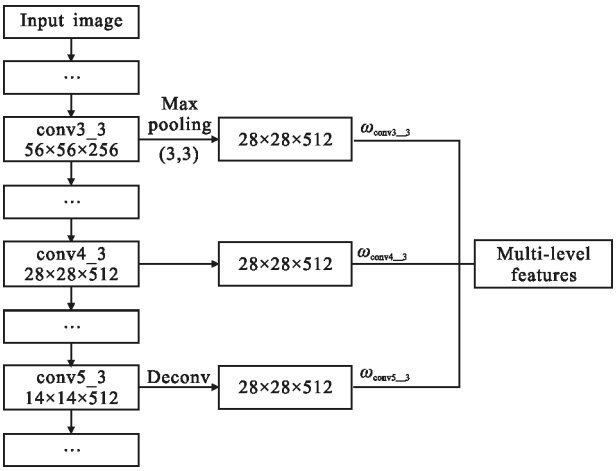


图 2 多层 CNN 特征的融合

Fig.2 Fusion of multi-level CNN features

1.4 训练分类模型用于视觉场景识别

文中把场景外观剧烈变化下的场景识别看作二分类问题,设计一个适合训练特征差异图的分类模型。

特征差异图包含较多的语义信息,因此,分类模型的结构可以相对简单。如图 3 所示,该分类模型包含 1 个卷积阶段和 4 个全连接阶段,仅有的 1 个卷积阶段包含 1 个卷积层和 1 个池化层,前 3 个全连接阶段包含 1 个全连接层和 1 个 ReLU 层,最后 1个全连接阶段只包含一个全连接输出层。训练中采用 xavier[22]算法进行权重初始化能取得更好的效果,与传统的 Gaussian 初始化不同,该算法可以使得前向传播和后向传播时每一层输入和输出的方差尽量相等,权重的初始值满足以下的分布:

$$W\sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}},\frac{\sqrt{6}}{\sqrt{n_j+n_{j+1}}}\right] \tag{6}$$

式中:n_j 为输入神经元的个数;n_{j+1} 为输出神经元的个数。

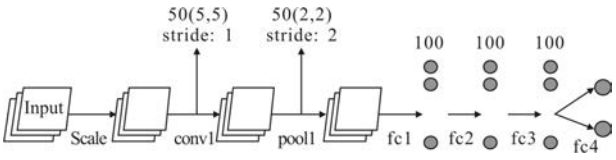


图 3 提出的分类模型结构

Fig.3 Architecture of the proposed classification model

2 实 验

首先介绍了文中采用的数据集,然后展示了由单层 CNN 特征构建的特征差异图和融合多层 CNN 特征构建的特征差异图的识别性能,同时也对提出的分类模型与现有的两个常用分类模型进行对比,最后与基于 CNN 特征的距离和阈值的方法进行对比。利用 caffe[23]进行实现,显卡为 8G 内存的 GTX1070。

2.1 数据集

文中所用的原始数据集为 Tokyo 24/7[24]。该数据集包含 375 组图片,每组的 3 张图片均为同一个地方在不同时刻从不同角度拍摄得到的(如图 4(a)~(c)所示)。为了获得更多的发生感知偏差和感知变异的图片,对原始数据集采取如下措施:(1)从来自同一场景的 3 张图片中任取两张可以获得 3 组图片对;

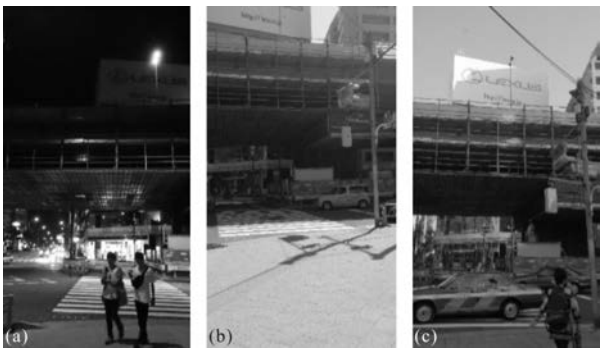


图 4 (a)~(c) 为同一场景在白天,傍晚和晚上采集的感知变异图片;(d)~(e) 为不同场景发生感知偏差的图片

Fig.4 Perceptually different images of the same place in (a)–(c) are captured at different times of day: daytime, sunset and night; aliased image pairs of different places are shown in (d)–(e)

(2)对来自不同场景的图片从光照,视角,内容等方面入手找到外观极其相似的图片对(如图 4(d)~(e)所示)。最终一共得到 1 125 对来自同一场景的发生感知变异的图片对和 1 500 对来自不同场景的发生感知偏差的图片对,并且从中挑选 25%作为测试集。

2.2 单层特征构建的特征差异图的性能

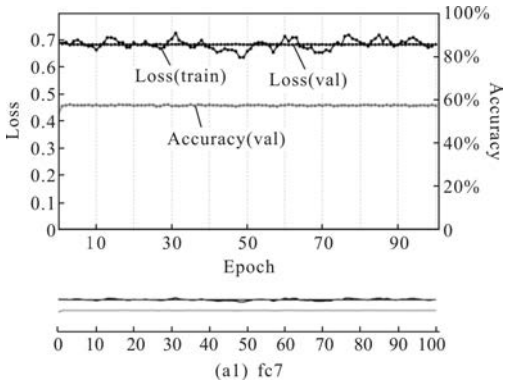
为了更好地融合多层 CNN 特征,利用单层特征构建的特征差异图进行训练并分析其识别性能。表2列出不同层(最浅的第一个卷积阶段忽略)在测试集上的精确度(precision)和准确率(accuracy)。图 5 给出了 fc7 和 conv2_2,conv5_3 和 conv3_3, conv4_3 和 pool4 的训练结果做直观的对比。

表 2 不同层在测试集上的精确度和准确度

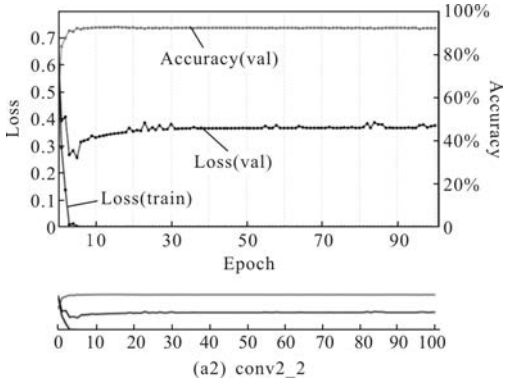
Tab.2 Precision and accuracy on the test set for different layers

Layer	Precision	Accuracy
conv2_1	0.686 12	0.593 33
conv2_2	0.714 82	0.736 67
pool2	0.723 89	0.726 67
conv3_1	0.864 82	0.873 33
conv3_2	0.856 21	0.863 33
conv3_3	0.888 16	0.893 33
pool3	0.883 82	0.883 33
conv4_1	0.886 08	0.906 67
conv4_2	0.890 32	0.903 33
conv4_3	0.861 64	0.883 33
pool4	0.888 89	0.896 67
conv5_1	0.896 10	0.906 67
conv5_2	0.918 37	0.910 00
conv5_3	0.908 50	0.916 67
pool5	0.904 46	0.923 33
fc6	0.783 12	0.793 33

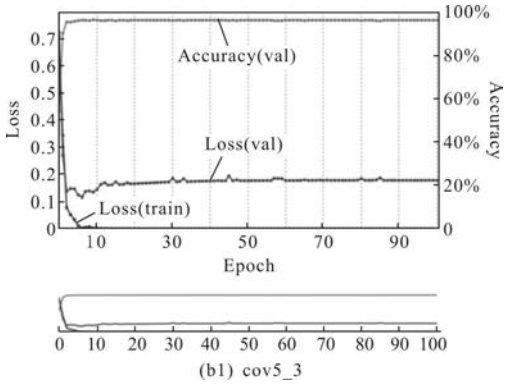
从实验结果可以看出:第一,与分类任务不同,对于场景识别来说, 卷积阶段的效果要明显好于全连接层;第二,中高层特征的性能显著优于浅层次的特征,验证了在场景识别任务中深层特征的重要性;



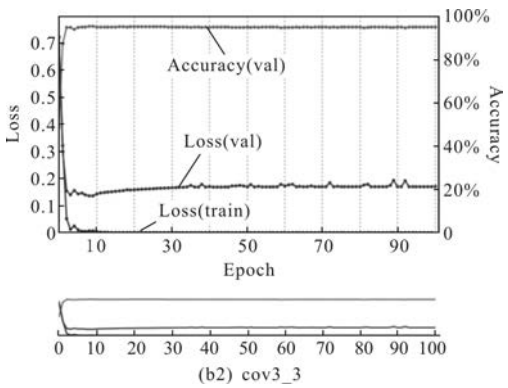
(a1) fc7



(a2) conv2_2



(b1) cov5_3



(b2) cov3_3

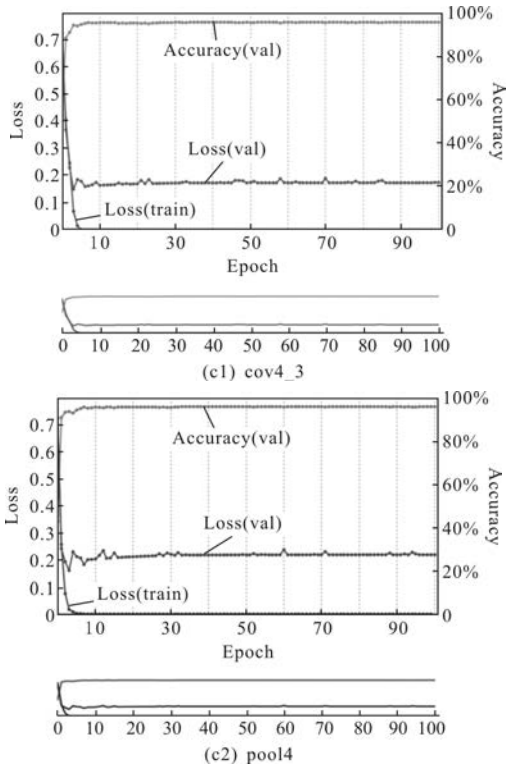


图 5 单层特征训练结果

Fig.5 Training result of mono-level features

第三,后三个卷积阶段均具有较高的性能,说明了融合中高层特征才能得到对光照、视角等更鲁棒的特征;第四,同一个卷积阶段中卷积层和池化层的性能相差不大,但是因为池化层得到的特征维数较低,所以从计算效率来说更占优势。

2.3 多层特征构建的特征差异图的性能

为了进一步说明融合多层 CNN 特征的必要性,首先利用反卷积对 pool3,pool4 和 pool5 三层的 feature maps 进行可视化,如图 6 所示,可以看出pool3(中等层次)提取了图片较多的边沿轮廓信息,而 pool5(高层次)则包含了更多的语义信息,融合这些不同层的特征才能得到对光照,视角等更鲁棒的特征。

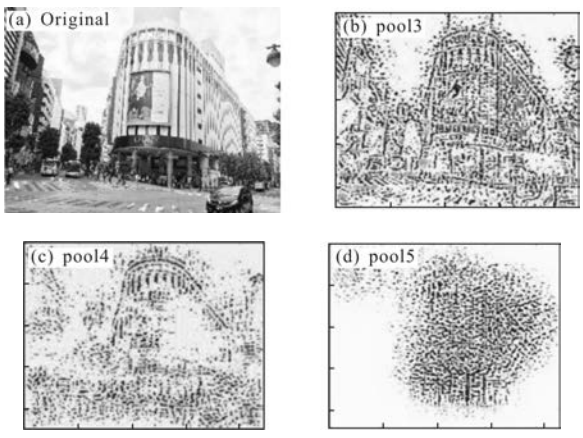


图 6 反卷积可视化结果

Fig.6 Visualization results using deconvolution

通过 2.2 节中对单层 CNN 特征构建的特征差异图的性能分析,同时也考虑到计算效率,采用了三种方式的融合:(1) pool5+conv5_3+conv5_2 (测试集上准确率最高的三层);(2) pool5+conv4_1+conv3_3 (均为所属卷积阶段中准确率最高的层);(3) pool3+pool4+pool5 (均为所属卷积阶段中准确率较高的池化层)。以 pool3+pool4+pool5 为例对融合过程进行说明,对 pool3 首先进行 max pooling 得到 14×14×256 的 feature maps,然后进行 3×3×512 的卷积转化为 14×14×512 的 feature maps;对 pool5 进行反向 pooling 和三个反卷积得到 14×14×512 的 feature maps,pool4 保持不变,然后 ω_{pool3},ω_{pool4},ω_{pool5} 均设为 1/3,最终将得到的 14×14×512 的向量按照 1.3 中的方法转化为 196×512 的特征差异图即为融合 pool3,pool4 和 pool5 构建的多层次特征差异图,图 7 为从多层次特征差异图中任意抽选的两张,d 和 s 分别表示其原始图片来自不同场景和相同场景。此外,需要说明的是,在多次实验中发现各层的权重趋于相等时识别效果较好,因此,融合中各层的权重均设为 1/3。图 8 展示了上述三种融合方式的训练结果。表 3 列出了其在测试集上的精确度和准确率。

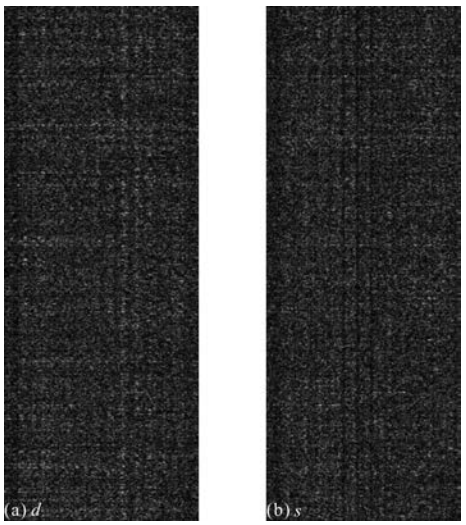


图 7 pool3+pool4+pool5 的多层次特征差异图示例

Fig.7 Example of multi-level feature difference map for pool3+pool4+pool5

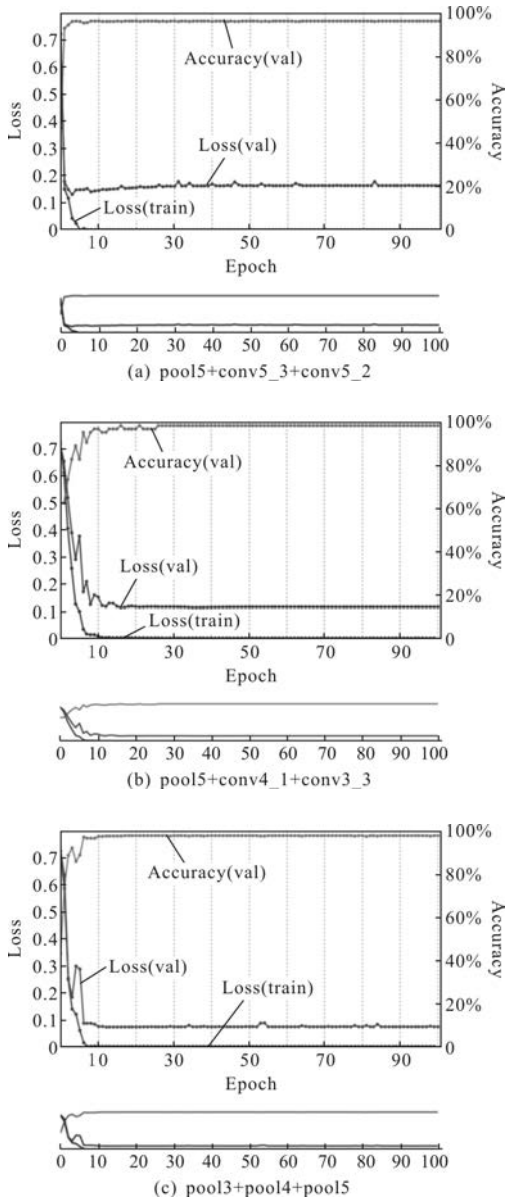


图 8 多层次特征差异图训练结果

Fig.8 Training results of multi-level feature difference map

表 3 多层次特征差异图在测试集上的精确度和准确率

Tab.3 Precision and accuracy on the test set for multi-level feature difference map

Layer	Precision	Accuracy
pool5+conv5_3+conv5_2	0.918 72	0.926 67
pool5+conv4_1+conv3_3	0.937 56	0.936 67
pool3+pool4+pool5	0.948 13	0.943 33

从实验结果中可以看出:(1) 对比表 2 和图 5 可以明显看出多层次特征差异图比单层次特征差异图具有更高的识别准确率，从而说明了融合后的特征更能有效地克服感知偏差和感知变异;(2) 后两种融合方式优于第一种融合方式，表明对于视觉场景识别任务来说，中等层次的边缘轮廓信息和高层次的语义信息同样重要;(3) 虽然后两种融合方式效果相差不大,但从计算效率来说,均为池化层的第三种融合方式更有优势。

2.4 分类模型的有效性

现有的分类模型可能并不适合训练特征差异图,为了说明文中所提出的分类模型的有效性,采用 pool5 构建的特征差异图对常见的两个分类模型 AlexNet 和 GoogLeNet 进行训练得到识别结果。实验结果如图 9 所示,两者的准确率都较差,说明了文中提出的适当增加全连接层减少卷积层的分类模型更适合训练特征差异图。

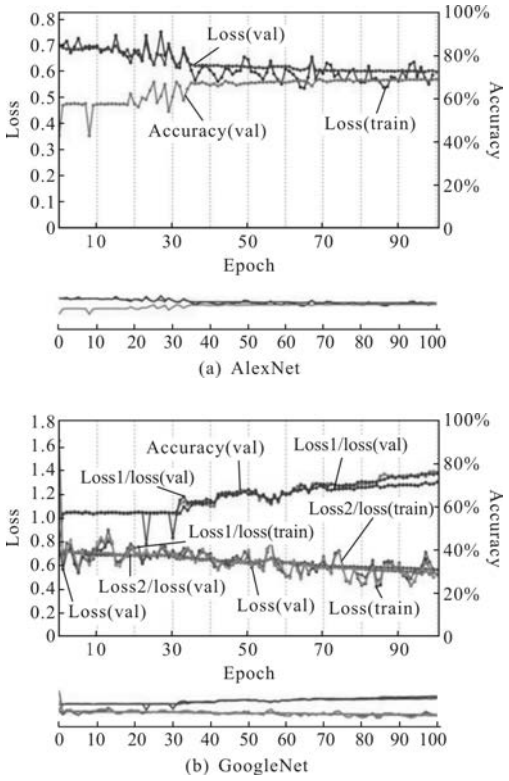


图 9 AlexNet 和 GoogleNet 的训练结果

Fig.9 Training results on AlexNet and GoogleNet

2.5 与基于 CNN 特征的距离和阈值相关算法的对比

现有的利用 CNN 特征的距离并设置阈值来衡量两幅图像相似性的相关算法，当场景外观剧烈变化时,很难选择合适的阈值进行准确的识别,效果较差。下面对比了在发生感知偏差和感知变异的数据

集 Tokyo 24/7 上文中提出的方法与三种基于 CNN 特征的距离和阈值方法的识别效果。P-R(Precision-recall)曲线和平均精确度(Average precision)被用来评价识别结果。

参考文献[15]是一种直接利用 CNN 特征的欧式距离并设置阈值来衡量图片相似性的方法(方法 1)。文中特征向量 $F_k(I)$均采用公式(7)中的 l_2 范数进行标准化。其 P-R 曲线如图 10 所示,此外,为了更直观地和文中方法的精确度进行对比,表 4 列出了其不同层在测试集上的平均精确度。

$$F_{\text{norm-k}}(I)=\left(\frac{x_{k1}}{\sqrt{\sum_{i=1}^{d}x_{ki}^{2}}},\cdots\frac{x_{kd}}{\sqrt{\sum_{i=1}^{d}x_{ki}^{2}}}\right) \tag{7}$$

式中:d 为特征向量的维数。

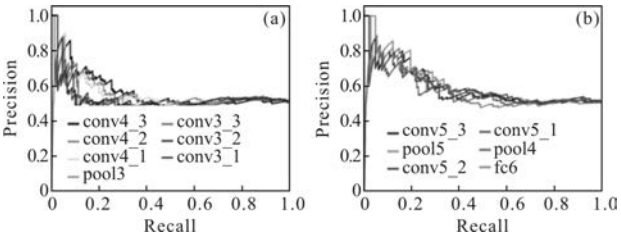


图 10 方法 1 不同层的 P-R 曲线

Fig.10 P-R curves for different layers of the method 1

表 4 方法 1 不同层的平均精确度

Tab.4 Average precision for different layers of the method 1

Layer	Average precision
conv3_1	0.530 94
conv3_2	0.529 74
conv3_3	0.518 54
pool3	0.522 03
conv4_1	0.545 27
conv4_2	0.553 00
conv4_3	0.554 68
pool4	0.564 65
conv5_1	0.577 40
conv5_2	0.596 21
conv5_3	0.590 79
pool5	0.601 27
fc6	0.591 59

参考文献[12]为了克服感知偏差,对由 CNN 特征的距离得到的混淆矩阵使用了两个连续滤波 (方法 2)。参考文献[14]同样融合了多层 CNN 特征来考察图片之间的相似性,与文中方法不同的是,其通过对低层次 feature maps 求熵,高层次 feature maps 求平均来融合不同层的特征(方法 3)。两种方法的 P-R 曲线如图 11 所示,表 5 列出了在测试集上的平均精确度,其中方法 2 采用 Overfeat 模型的第 10 层(文中效果较好的层)进行特征提取。

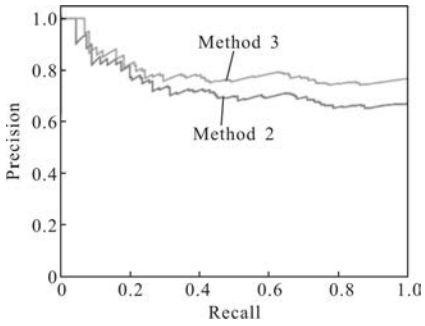


图 11 其他两种方法的 P-R 曲线

Fig.11 P-R curves of the other two methods

表 5 其他两种方法的平均精确度

Tab.5 Average precision of the other two methods

Method	Average precision
Method 2	0.732 59
Method 3	0.795 90

从实验结果可以看出:(1) 当图片发生严重感知偏差和感知变异时,利用 CNN 特征的距离衡量图片相似性的方法不能很好地选择合适的阈值来判断两幅图片是否来自同一场景,识别准确率较低;(2) 方法 3 的效果要明显优于使用单层特征的方法,说明了融合多层 CNN 特征才能得到对光照,视角等更鲁棒的特征;(3) 对比文中基于多层次特征差异图的方法,说明了融合多层 CNN 特征构建的多层次特征差异图能很好地表征图像之间的差异,当场景外观发生剧烈变化时仍能取得较高的识别准确率。

3 结 论

针对场景外观剧烈变化引起的感知偏差和感知变异,文中提出了一种新的融合多层 CNN 特征构建特征差异图的视觉场景识别方法,得出如下结论:

(1) CNN 不同层特征对光照,视角等的鲁棒性不同,结合不同层尤其是中高层的特征能够获得更好的识别性能;

(2) 当发生感知偏差和感知变异时,特征差异图不仅能很好地表征图像之间的差异，也避免了设置阈值，当场景外观发生剧烈变化时仍能取得很高的准确率；

此外，从原始数据中获得更多的感知偏差和感知变异的图片，以及设计适合特征差异图训练的分类模型为所提出方法提供了重要的保证。

参考文献：

[1] Lowry S, Sünderhauf N, Newman P, et al. Visual place recognition: A survey [J]. ***IEEE Transactions on Robotics***, 2016, 32(1): 1-19.

[2] Lowe D G. Object recognition from local scale-invariant features [C]//The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, 2: 1150-1157.

[3] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF) [J]. ***Computer Vision and Image Understanding***, 2008, 110(3): 346-359.

[4] Cummins M, Newman P M. Appearance-only SLAM at large scale with FAB-MAP 2.0 [J]. ***International Journal of Robotics Research***, 2011, 30(9): 1100-1123.

[5] Angeli A, Filliat D, Doncieux S, et al. Fast and incremental method for loop-closure detection using bags of visual words [J]. ***IEEE Transactions on Robotic***s, 2008, 24(5): 1027-1037.

[6] Nister D, Stewenius H. Scalable recognition with a vocabulary tree [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, 2: 2161-2168.

[7] Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition[J]. ***Progress in Brain Research***, 2006, 155: 23-36.

[8] Blaer P, Allen P. Topological mobile robot localization using fast vision techniques [C]//IEEE International Conference on Robotics and Automation, 2002, 1: 1031-1036.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems, 2012: 1097-1105.

[10] Babenko A, Slesarev A, Chigorin A, et al. Neural codes for image retrieval [C]//European Conference on Computer Vision. Springer International Publishing, 2014: 584-599.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

[12] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)

[13] Bao Xuejing, Dai Shijie, Guo Cheng, et al. Nonlinear distortion image correction from confocal microscope based on interpolation [J]. ***Infrared and Laser Engineering***, 2017, 46(11): 1103006. (in Chinese)

[14] Li Q, Li K, You X, et al. Place recognition based on deep feature and adaptive weighting of similarity matrix [J]. ***Neurocomputing***, 2016, 199: 114-127.

[15] Hou Y, Zhang H, Zhou S. Convolutional neural network-based image representation for visual loop closure detection [C]//IEEE International Conference on Information and Automation, 2015: 2238-2245.

[16] Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database [C]//Advances in Neural Information Processing Systems, 2014: 487-495.

[17] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[18] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010: 3304-3311.

[19] Sünderhauf N, Shirazi S, Dayoub F, et al. On the performance of convnet features for place recognition [C]// IEEE International Conference on Intelligent Robots and Systems, 2015: 4297-4304.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. ***Computer Vision and Pattern Recognition***, 2014, arXiv preprint arXiv:1409.1556.

[21] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks [C]//European Conference on Computer Vision, 2014: 818-833.

[22] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010: 249-256.

[23] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding [C]//Proceedings of the 22nd ACM international conference on Multimedia, 2014, arXiV preprint arxiv: 1408.5093.

[24] Torii A, Arandjelovic R, Sivic J, et al. 24/7 place recognition by view synthesis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2667665.