

基于高维特征表示的交通场景识别

刘文华, 李浥东, 王涛, 邬俊, 金一

(北京交通大学计算机与信息技术学院, 北京 100044)

摘要: 随着智能交通的发展, 快速、精确识别交通场景成为亟待解决的重要问题。目前已有许多识别方法可以提高交通场景的识别效果, 但这些算法无法提取视觉概念的交通语义特征, 导致识别精度低下。为此, 设计了一种提取高维场景语义特征和结构信息的识别算法, 以提高识别精度。为减少图像高维与低维特征表示之间的“语义鸿沟”, 首先构建了一个场景类的语义描述系统, 然后通过最小化损失(element-wise logistic loss)函数训练多标签分类网络, 获取交通场景图像的高维特征表示, 最后在4个大规模场景识别数据集上进行验证, 实验结果显示, 新算法在识别性能上优于其他的方法。

关键词: 场景识别; 卷积神经网络; 高维特征; 低维特征

中图分类号: TP30

文献标识码: A

doi: 10.11959/j.issn.2096-6652.201943

Transportation scene recognition based on high level feature representation

LIU Wenhua, LI Yidong, WANG Tao, WU Jun, JIN Yi

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Abstract: With the development of intelligent transportation, it has become an urgent problem to quickly and accurately recognize complex traffic scene. In recent years, a large number of scene recognition methods have been proposed to improve the effectiveness of traffic scene recognition, however, most of these algorithms cannot extract the semantic characteristics of the concept of vision, leading to the low recognition accuracy in traffic scenes. Therefore, a novel traffic scene recognition algorithm which extracts the high-level semantic and structural information for improving the accuracy was proposed. A system to discover semantically meaningful descriptions of the scene classes to reduce the “semantic gap” between the high level and the low-level feature representation was built. Then, the multi-label network was trained by minimizing loss function (namely, element-wise logistic loss) to obtain the high-level semantic representation of traffic scene images. Finally, experiments on four large-scale scene recognition datasets show that the proposed algorithm considerably outperforms other state-of-the-art methods.

Key words: scene recognition, CNN, high-level feature, low-level feature

1 引言

场景识别已被广泛应用于图像处理任务中, 如图像检索、行为检测和目标识别。场景识别问题本质上是图像语义分类问题, 通过使用场景特征向量

表示图像, 将场景图像分为站台、轨道、变道等不同类别的场景。特征表示和分类方法是影响场景识别效果的关键。

特征表示是场景识别处理任务的第一步, 特征表示的效果直接影响识别精度, 如何提取具有判别

收稿日期: 2019-10-14; 修回日期: 2019-11-26

通信作者: 刘文华, 166112077@bjtu.edu.cn

基金项目: 国家自然科学基金资助项目(No.61672088)

Foundation Item: The National Natural Science Foundation of China(No.61672088)

性的特征成为新的研究热点。图像特征中包含的图像信息越多,识别效果越好。目前,许多用于提取图像特征的算法被陆续提出,较为经典的有:基于梯度的 GIST (generalized search tree) 算子^[1],该算子专门用于描述图像空间特征,以估计全局空间特征;基于纹理的尺度不变特征转换 (scale-invariant feature transform, SIFT) 算子^[2],该算子检测尺度空间中的特征,并识别关键点的位置。这两种方法获得的特征是图像的低维统计信息,将低维特征直接进行映射,导致对图像的识别效果差。为解决低维表示直接映射引起的“语义鸿沟”问题, Li 等人^[3]提出了一种高层语义表示方法——对象库 (OB),它由许多对象特征图构成,对图像对象的语义和空间信息进行编码,图像被表示为通用对象的特征图。图像的 GIST 特征图、SPM 表示直方图和 OB 对象特征图具体如图 1 所示。OB 算法表示图像的效果明显优于前两种算法,语义信息更丰富,但 OB 提取的特征中不包含描述性的图像信息。

近年来,深度神经网络给计算机视觉领域带来了革命性的变革,极大地推动了场景识别技术的发展。深度神经网络中场景识别任务是从已标注的训练图像集中学习语义模型,建立低层视觉特征到高层语义概念之间的映射。基于已有研究方法,本文提出了一种新的图像语义表示方法,以减小高层识别任务与低层特征之间的“语义鸿沟”。本文所提算法将提取图像特征过程视为多标签分类,每个语义特征对应训练网络,得到分类标签,并在 4 个经典场景数据集上进行了验证。

2 特征表示

特征表示有低维特征表示和高维特征表示,具体如下。

低维特征表示方法是指从图像纹理、颜色、形状等方面来表示图像^[4-6]。这些算法可解释性强且时间复杂度低,但其表示性能较差。随后,中层特征表示方法被提出,其中被广泛使用的方法有计算图像每个子区域局部特征直方图的 SIFT 空间金字塔算法^[5]、词袋 (bag of words) 模型^[6]、提取图像结构特征的视觉描述符算法^[7]等。研究发现,图像中低维特征的直接映射可能会引起更大的“语义鸿沟”。为此, Fang H^[8]、Farhadi A^[9]、Li F F^[10]等人提出了从低维特征表示中学习高维语义的学习模型,获得了优异的识别性能,在 Sports 数据集上取得的识别准确率分别为 84.4%、84.92%和 85.7%。Felzenszwalb P F^[11]提出了一种基于多种特征的混合特征提取方法,该方法在 MIT-Indoor 数据集上的识别精度达到了 43.1%。

高维特征表示方法是由低维特征组合而成的更结构化、更复杂的特征表示方法。比如,经典的 OB 算法首先利用目标检测器^[12-14]获得图像中包含的对象,然后编码对象的语义和空间信息。随后, Hauptmann^[15]使用对象到类 (O2C) 的距离构建场景识别模型,基于对象库的 O2C 距离的显式计算获得的图像特征更加抽象、复杂。现实世界中的对象具有层次结构概念,这导致基于对象的表示方法可能存在语义层次结构问题,无法同时识别同一张图中的“车”和“汽车”。为了解决语义层次结构问

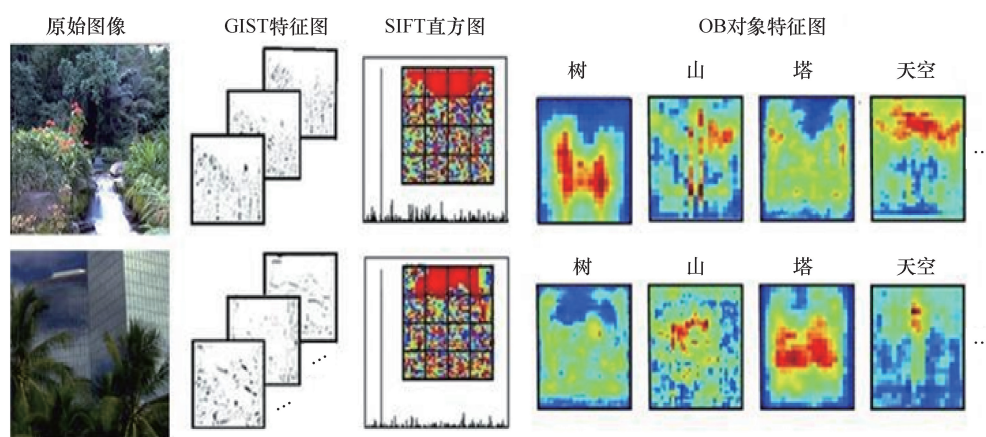


图1 图像特征结果

题,使用语义特征作为图像的代表方法,在没有训练图像的情况下,语义特征可用于识别对象类,使用语义特征识别对象类的这一过程被称为 zero-short learning。为了获取图像的全局和局部特征,Vogel 等人用图像局部语义与图像全局特征组合来表示图像。Hinton G E 等人^[16]定义了中级特征,设计了 6 组特征,并将其作为图像的中层语义描述符。实验表明,图像特征包含的信息越丰富,其识别准确率越高。

3 交通场景识别算法

本文定义了一种基于深度卷积神经网络(convolutional neural network, CNN)的图像特征提取模型,并利用该模型进行场景识别。卷积神经网络通过卷积核提取每层特征,前一层的输出作为后一层的输入,递归获取图像特征,卷积神经网络的深度越深,获得的特征越抽象,语义信息越丰富,越有利于学习识别算法。基于深度卷积神经网络的图像特征提取模型主要包含 3 步:首先通过在图像数据集 ImageNet^[17]上预训练模型 VGGNet (visual geometry group)^[18]来初始化网络参数,以提高网络的收敛速度;然后利用多标签 COCO (common objects in context) 数据集^[19]调整网络参数,构建多标签预测网络;最后对场景数据集进行特征预测。具体过程如图 2 所示。

3.1 字典设计

在预测图像特征之前,需构造一个特征词汇表,这里的特征应涵盖场景图像包含的对象,如图像可视对象、抽象对象或意象对象等。本文词汇表是基于文献[20-21]的思想进行构建的,首先利用多标签 CNN 模型获得图像对应的特征向量 $V_{aj}=[v_{1j},v_{2j},v_{3j},\cdots,v_{cj}]$,其中 v_{ij} 为图像 j 第 i 个特征的预测概率,然后判断图像集中所有图像的特征 i 的预测概率是否大于预设参数 γ ,若 $v_{ij} > \gamma$, $j \in \{\text{imageset}\}$, v_{ij} 就被加入词汇表 V 中。词汇表 V 构建完成之后,利用多标签 CNN 模型预测图像的特征概率 $[p_{1j},p_{2j},\cdots,p_{mj},p_{vj}]$,然后从中选择前 t 个最显著的特征,为每个图像构造一个固定长度的向量。

3.2 图像特征提取

随着对象数量的增加,图像中对象的层次问题更加显著。例如,对一张包含“公交车”的图像,OB 算法无法提取出“车”和“公交车”两个特征,因为 OB 算法设定对象向量是二进制的,当对象位于图像中时,图中对象对应的二进制向量值为 1,否则为 0。但从现实分析,“公交车”属于“车”的一种,包含“公交车”的图像应具有“车”的特征。由于目前无法提取具有层次特征的图像,为解决这一问题,现有方法^[22]采用多层次、多尺度特征融合策略,经过金字塔池化各级特征,由深到浅逐步融合,最后输入 soft max 分类器,输出“公交车”或者“车”(不并

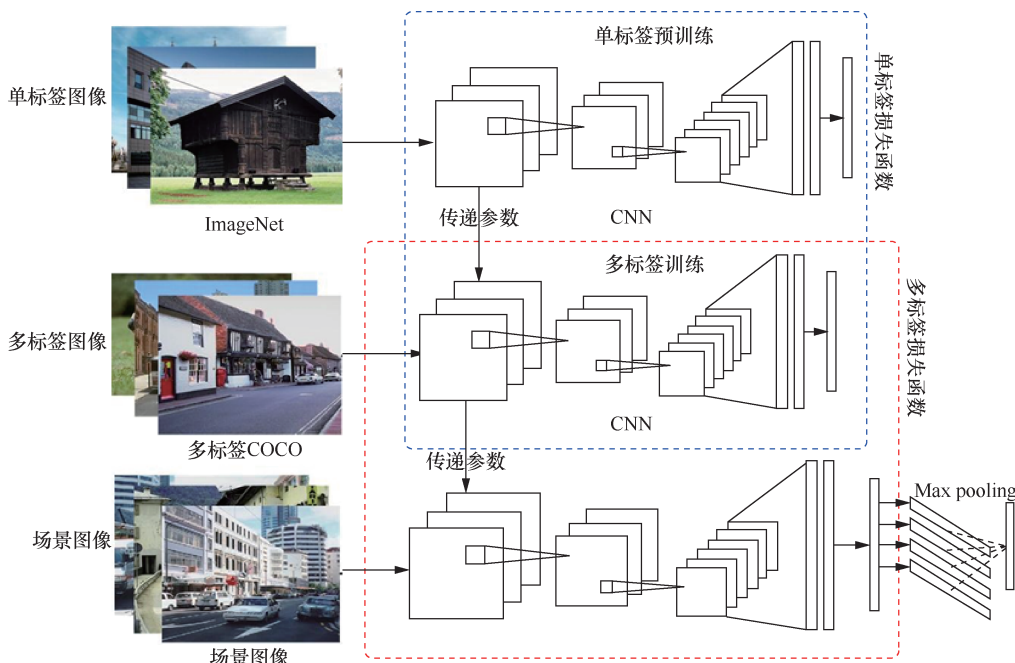


图 2 图像特征提取过程

存)。为解决同时预测层次类对象问题,本文提出一种基于词汇表的特征概率向量(特征集)表示方法。假设词汇表为 $V=[\text{car}, \text{bus}, \text{train}, \text{sedan}, \dots, \text{building}, \text{tower}, \text{pagoda}, \text{gallery}, \dots]$,本文算法旨在预测时得到对应于词汇表的特征概率向量 $V_p=[0.45, 0.89, 0.31, 0.28, \dots, 0.56, 0.37, 0.22, 0.14, \dots]$ 。图像特征向量中的每个特征具有不同的概率值,抽象程度越高的特征(对象越具体)对应的概率向量值越大。

为了增加特征的表述能力,本文引入多标签分类概念。多标签分类算法可获取多类特征,即 $\text{Input}(\text{image}) \rightarrow \text{Output}(\text{label}(\text{bus}), \text{label}(\text{tree}), \dots, \text{label}(\text{building}))$,但这些特征都是名词性特征,描述性差。为提高图像特征的描述性,本文提出一种基于深度模型的多标签分类算法,利用该算法提取多类型特征,即 $\text{Input}(\text{image}) \rightarrow \text{Output}(\text{label}(\text{bus}), \text{label}(\text{red}), \dots, \text{label}(\text{drive}))$ 。预测的特征包含名词(场景标签)、形容词(场景颜色)、动词(场景特征)等,这是与传统多标签分类算法最大的不同点。为提高算法的收敛速度,首先在ImageNet数据集上预训练VGG16,获取网络初始参数,修改网络损失函数为 c -way multi-logistic,并在COCO数据集上调整网络参数,将特征向量的大小 c 设置为256。在预测图像特征时,将预测模型的 c 值设置为特征集的大小, c 个单元的multi-logistic输入替换最后一个全连通层的输出,这样通过 c -way得到概率分数在多类标签上呈现出的一个概率分布,即图像特征的概率分布。

损失函数的定义:假设训练集中有 N 个样本和 c 个与图像中特征相关的多标签,第 i 幅图像的多标签向量为 $y_i=[y_{i1}, y_{i2}, y_{i3}, \dots, y_{ic}] \in \{0, 1\}$,当第 j 个特征包含在第 i 幅图像中时 $y_{ij}=1$,否则, $y_{ij}=0$ 。将向量 $p_i=[p_{i1}, p_{i2}, p_{i3}, \dots, p_{ic}]$ 作为预测概率向量与 $y_i=[y_{i1}, y_{i2}, y_{i3}, \dots, y_{ic}]$ 对应。模型的损失函数为:

$$J = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^c \log(1 + \exp(y_{ij} p_{ij})) \quad (1)$$

3.3 图像识别模型

场景识别模型^[23-26]将提取的特征向量作为输入得到识别结果,即 $f(x)=c_{\text{label}}$ 。输入 $x=[x_1^T; x_2^T; \dots; x_N^T] \in R^{N \times D}$, R 为 $N \times D$ 维矩阵, D 表示特征词典 V 的维数, N 为场景图像的个数, $C=(C_1, C_2, \dots, C_N) \in \{0, 1\}^N$,表示 N 幅图像场景标签。本文采用两种类型的函数作为 $f(\bullet)$ 。

① $f(x)=\text{sgn}()$, $\text{sgn}(\bullet)$ 是符号函数,具体定义由式(2)给出。

$$f(x)=\text{sgn}(g(x))=\begin{cases} \text{label}=C_1, & g(x) \geq 0 \\ \text{label}=C_2, & g(x) < 0 \\ \text{rejected}, & \text{其他} \end{cases} \quad (2)$$

其中, $g(x)=wx+b$, $g(x)$ 是一组将 w 作为法向量的超平面。 C_1 表示一个场景类, C_2 表示其余数据集中的剩余场景。最大化两类之间的分类区间是选择最优分类超平面的关键条件,即:

$$\max \frac{1}{\|w\|} |g(x)| \quad (3)$$

其中, $\|w\|$ 是归一化的权重,为了简化计算,将最大化分类间隔 δ 转为求解最小化 $\frac{1}{2}\|w\|^2$ 。

② $f(x)=h_\beta$, $h_\beta \triangleq \arg \max c \in \{0, 1\}^{x^\beta}$, $\beta=(\beta_1, \beta_2, \dots, \beta_D) \in R^D$ 表示特征向量的参数,求解 β 的目标函数由式(4)给出。

$$\min_{\beta \in R^D} \theta R(\beta) + \frac{1}{N} \sum_{i=1}^N L(\beta; x_i; y_i) \quad (4)$$

其中, $L(\beta; x_i; y_i)$ 一般选用的是非负函数或凸函数; $R(\beta)$ 是正则化函数,用来避免过度拟合; $\theta \in R$ 是正则化系数,由交叉验证获得。本算法中采用 $\log$ 函数作为 $L(\bullet)$ 函数。

$$L = \log(1/P(c_i | x_i, \beta)) \quad (5)$$

其中, $P(c | x, \beta) = \frac{1}{N} \exp\left(\frac{1}{2} c(x \cdot \beta)\right)$ 。

4 实验设计与验证

4.1 数据集

本次实验选取4个经典场景数据集Sport^[27]、Indoor^[28]、Outdoor^[29]、15 Scene^[30]进行性能测试,具体介绍见表1。Outdoor、15 Scene数据集中包含丰富的复杂交通场景图像,如图3所示。

表1 场景数据集描述

数据集	描述
Sport	数据集包含8个运动场景
Indoor	数据集共15 620张图片,含67类室内场景,每类至少包含100张图
Outdoor	数据集含有8类室外场景,共2 688张图
15 Scene	数据集是基于文献[11]创建的,共包含15种类别的自然场景

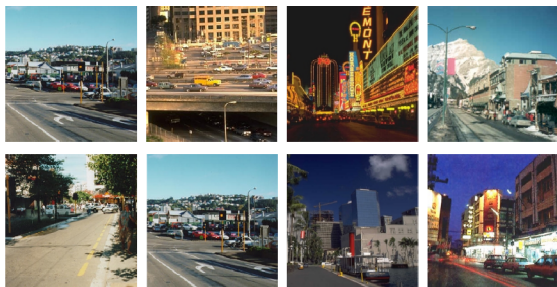


图 3 交通场景图像

4.2 实验参数设置

本算法设计的网络共包含 16 层，其中包括 13 个卷积层和 3 个全连接层。模型中所有卷积滤波器的大小被设置为 3×3 ，卷积滤波器的接收域被设为 7×7 ，这可以代替较大的滤波器。为提高模型的收敛速度，利用文献[31]中的网络参数初始化模型中的最后两个全连接层的参数 w 。本算法在 ImageNet 数据集上进行预训练时，将全连接层 f6 和 f7 的学习率均设为 0.001，全连通层 f8 的学习率设为 0.01，动量和下降速率均设为 0.9 和 0.5。为避免陷入局部最优，在计算连接层 w_{ij} 的参数值时，按预设比例降低各层的学习率，具体网络结构如图 4 所示（解释了 13 个卷积层和 3 个全连接层的设计结构）。

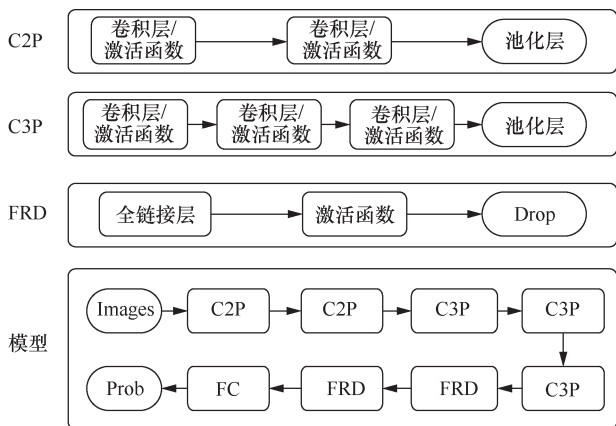


图 4 网络结构

4.3 对比实验

为了验证算法的有效性，将本文所提算法与两大类图像特征提取算法进行比较：低维语义表示方法（GIST^[32]、SIFT^[33]和 CENTIRST^[34]）和高维语义表示方法（OB、KCL^[35]），对比结果见表 2（Attributes_{Finetune} 表示本模型经过 Finetune 提取的图像特征）。

实验结果表明，高维语义表示算法比低维语义表示算法有更好的识别性能，说明高维特征包含判

别性的语义信息，消除了场景识别任务和特征表示之间的“语义鸿沟”。本文所提算法的优势为在识别复杂场景图像时更为突出，例如，对包含杂乱图像的 Outdoor 数据集的识别精度提升更高。

表 2 场景识别准确率对比

数据集	Sport	Indoor	Outdoor	15 Scene
GIST	82.6%	5.7%	81.9%	73.3%
SIFT	82.8%	44.2%	15.67%	82.4%
CENTIRST	86.2%	31.9%	89.57%	83.9%
OB	77.5%	33.3%	88.12%	82.0%
KCL	86.0%	32.4%	88.8%	88.8%
Attributes _{Finetune}	96.23%	68.32%	98.83%	91.92%

为验证网络结构的有效性，选取经典的 CaffeNet、AlexNet、VGGNet16 和 Places-CNN 进行网络结构对比，实验结果见表 3。

表 3 不同网络结构的识别性能对比

网络结构	识别准确率			
	Sport	Indoor	Outdoor	15 Scene
CaffeNet	94.91%	60.05%	94.54%	88.62%
AlexNet	94.06%	58.15%	93.42%	86.84%
VGGNet16	96.21%	64.13%	92.30%	91.23%
Places-CNN	94.12%	68.24%	95.23%	90.19%
Soft _{attribute}	93.87%	66.43%	98.71%	88.48%
Attributes _{Finetune}	96.23%	68.32%	98.83%	91.92%

从表 3 可以看出，CaffeNet 和 AlexNet 的识别准确率明显低于本文所提模型，VGGNet16 和 Places-CNN 的性能优于前两种算法，因为 VGGNet16 具有更深层次的网络结构，而 Places-CNN 是在特殊的场景数据集上训练的。从实验结果可知，网络结构层次越深，特征的抽象程度越高，识别效果越好。为验证损失函数对本文所提模型的影响，分别将 softmax、element-wise logistic loss 作为损失函数，即表 3 中的 Soft_{attribute}、Attributes_{Finetune}。

综上所述，本文所提模型所提取的特征可以承载更多的语义信息，将 element-wise logistic loss 作为损失函数可提升模型的识别效果。

模型的性能随调优次数的增加而变化，实验对迭代过程中保存的中间模型的识别准确率进行了比较，其结果如图 5 所示。

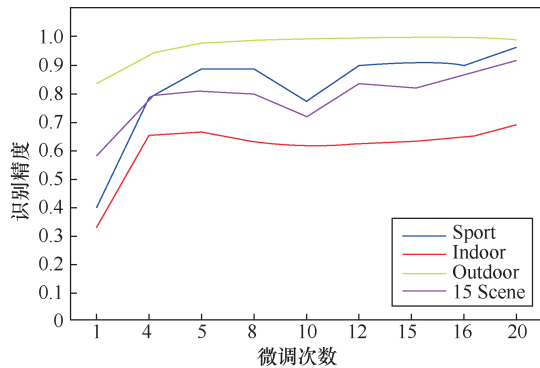


图5 识别精度随微调次数的变化趋势

从图5中可以看出,使用梯度搜索方法求解参数时,随着迭代次数的增加,模型的识别准确率总体呈上升趋势,有微小波动(可忽略不计)。实验表明,每隔20个时间间隔保存的模型性能良好,避免了过度拟合,在4个数据集上均获得了较高的识别准确率。

以往大量实验表明^[36-37]:场景识别算法性能与训练样本个数密切相关。在实验中,本文使用不同大小的训练数据对模型进行微调,并利用不同大小的特征训练分类器来测试其鲁棒性,训练范围为训练数据的50%、70%、100%,其中,将GIST提取的图像特征定义为 $Feature_{GIST}$ 。表4总结了训练数据大小对算法性能的影响。

表4 训练数据大小对算法性能的影响

算法	训练数据	Sport	Indoor	Outdoor	15 Scene
$Attributes_{Finetune}$	50%	85.32%	46.07%	92.61%	14.53%
	70%	89.73%	57.72%	93.68%	89.92%
	100%	96.23%	68.32%	98.83%	91.92%
$Feature_{GIST}$	50%	59.95%	4.89%	75.68%	58.93%
	70%	67.59%	5.69%	81.96%	67.68%
	100%	68.86%	5.92%	83.53%	69.27%

从表4可知,训练数据的比例从50%增加到100%时,本文所提算法的识别性能得到了很大的改善,当所有的训练数据都用于训练模型时,识别性能达到最优。然而,当GIST特征用于训练识别算法时,训练数据的大小对其性能影响较小,这也从另一个角度证明了GIST算子的局限性。在训练数据比例为50%的情况下,本文所提算法的识别准确率比GIST在训练数据比例为100%时更高,这充分证明了本文所提出的语义特征表示包含更多的判别信息,这些信息被隐藏在较低维的“信息”特征空间中。

不同的分类算法对场景识别的精度和效率都有影响,将提取的语义特征分别作为分类器线性支持向量机(linear SVM)和线性二叉树(linear binary tree)的输入,并进行对比,见表5。从表5可以看出,线性支持向量机在Indoor和15 Scene数据集上的分类效果优于线性二叉树算法,在另两个数据集上的分类效果却低于线性二叉树算法。这说明,针对不同的场景应选用不同的分类算法进行识别。

表5 不同分类器的识别性能对比

分类算法	识别准确率	
	线性支持向量机	线性二叉树
Sport	92.95%	96.23%
Indoor	73.38%	68.72%
Outdoor	93.38%	98.43%
15 Scene	93.86%	91.22%

5 结束语

本文所提算法提取的图像特征不仅有效地缩小了交通场景识别(高维视觉任务)与低维表示之间的“语义鸿沟”,而且解决了语义层次的问题。从网络结构上看,本模型是在VGGNet16网络结构基础上进行的改进,使用element-wise logistic函数作为损失函数。本文所提算法的技术贡献在于提出了一种新的特征表示方法,构建场景特征字典。实验结果表明,本文所提算法在4种场景数据集上的性能明显优于现有的算法。

参考文献:

- [1] AKER A, GAIZAUSKAS R. Generating image descriptions using dependency relational patterns[C]// Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. [S.l.:s.n.], 2010: 1250-1258.
- [2] AMRINE D E, WHITE B J, LARSON R L. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease[J]. Computers & Electronics in Agriculture, Computers and Electronics in Agriculture, 2014, 105: 9-19.
- [3] BELONGIE S, MALIK J, PUZICHA J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 24(4): 509-522.
- [4] BO L F, REN X F, FOX D. Hierarchical matching pursuit for image classification: architecture and fast algorithms[C]//NIPS, December 3-6, 2012, Lake Tahoe, USA. [S.l.:s.n.], 2012: 2115-2123.
- [5] CHERIYADAT A M. Unsupervised feature learning for aerial scene

- classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(1):439-451.
- [6] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]// 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), June 20-25, 2009, Miami, USA. Piscataway: IEEE Press, 2009: 248-255.
- [7] DIXIT M, RASIWASIA N, VASCONCELOS N. Adapted Gaussian models for image classification[C]// IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2011, Colorado Springs, USA. Piscataway: IEEE Press, 2011: 937-943.
- [8] FANG H, PLATT J C, ZITNICK C L. From captions to visual concepts and back[J]. Computer Science, 2015, 12(7): 1473-1482.
- [9] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: generating sentences from images[C]// European Conference on Computer Vision, September 5-11, 2010, Crete, Greece. Heidelberg: Springer, 2010: 15-29.
- [10] LI F F, PERONA P. A Bayesian hierarchical model for learning natural scene categories[C]// CVPR, June 20-26, 2005, San Diego, USA. Piscataway: IEEE Press, 2005: 524-531.
- [11] FELZENSZWALB P F, GIRSHICK R, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Software Engineering, 2014, 32(9): 1627-1645.
- [12] GAO S H, CHIA L T, TSANG I W. Multi-layer group sparse coding for concurrent image classification and annotation[C]// IEEE Conference on Computer Vision & Pattern Recognition, June 20-25, 2011, Colorado Springs, USA. Piscataway: IEEE Press, 2011: 2809-2816.
- [13] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feed forward neural networks[J]. Journal of Machine Learning Research, 2010, 9: 249-256.
- [14] HAN Y, LIU G Z. Efficient learning of sample-specific discriminative features for scene classification[J]. Signal Processing Letters IEEE, 2011, 18(11): 683-686.
- [15] HAUPTMAN N, YAN R, LIN W H, et al. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news[J]. IEEE Transactions on Multimedia, 2007, 9(5): 958-966.
- [16] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.
- [17] JIA Y Q, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]// ACM International Conference on Multimedia, October 23-27, 2006, Santa Barbara, USA. New York: ACM Press, 2014: 675-678.
- [18] KULKARNI G, PREMRAJ V, ORDONEZ V. Babytalk: understanding and generating simple image descriptions[C]// IEEE Conference on Computer Vision & Pattern Recognition, June 20-25, 2011, Colorado Springs, USA. Piscataway: IEEE Press, 2011: 1601-1608.
- [19] LAMPERT C H, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(3): 453-465.
- [20] LAPIN M, SCHIELE B, HEIN M. Scalable multitask representation learning for scene classification[C]// IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, USA. Piscataway: IEEE Press, 2014: 1434-1441.
- [21] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition, June 17-22, 2006, New York, USA. Piscataway: IEEE Press, 2006: 2169-2178.
- [22] 赵斐, 张文凯, 闫志远, 等. 基于多特征图金字塔融合深度网络的遥感图像语义分割[J]. 电子与信息学报, 2019, 41(10): 2525-2531.
- ZHAO F, ZHANG W K, YAN Z Y, et al. Multi-feature map pyramid fusion deep network for semantic segmentation on remote sensing data[J]. Journal of Electronics and Information Technology, 2019, 41(10): 2525-2531.
- [23] LI J L, LI F F. What, where and who? classifying events by scene and object recognition[C]// 2007 IEEE 11th International Conference on Computer Vision, October 14-21, 2007, Rio De Janeiro, Brazil. Piscataway: IEEE Press, 2007: 1-8.
- [24] LI J L, SU H, LIM Y W, et al. Object bank: an object-level image representation for high-level visual recognition[J]. International Journal of Computer Vision, 2014, 107(1): 20-39.
- [25] LI J L, SU H, XING E P, et al. Object bank: a high-level image representation for scene classification & semantic feature sparsification[C]// The 24th Annual Conference on Neural Information Processing Systems 2010, December 6-9, 2010, Vancouver, Canada. [S.l.: s.n.], 2010.
- [26] OLIVA AUDE, TORRALBA A. Modeling the shape of the scene: a holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2001, 42(3):145-175.
- [27] PANDEY M, LAZEBNIK S. Scene recognition and weakly supervised object localization with deformable part-based models[C]// IEEE International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. Piscataway: IEEE Press, 2011: 1307-1314.
- [28] PAVLOPOULOU C, YU S X. Indoor-outdoor classification with human accuracies: image or edge GIST[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 13-18, 2010, San Francisco, USA. Piscataway: IEEE Press, 2010: 41-47.
- [29] QUATTON I, TORRALBA A. Recognizing indoor scenes[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, USA. Piscataway: IEEE Press, 2009: 413-420.
- [30] RAMANA B V, BABU M S P. A critical study of selected classification algorithms for liver disease diagnosis[J]. International Journal of Database Management Systems, 2011, 3(2): 324-335.
- [31] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks

- for large-scale image recognition[J]. Computer Science, 2015.
- [32] TORRESANI L, SZUMMER M, FITZGIBBON A. Efficient object category recognition using classemes[C]// ECCV, September 5-11, 2010, Crete, Greece. Heidelberg: Springer, 2010: 776-789.
- [33] WU J X, REHG M. Centrist: a visual descriptor for scene categorization[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(8): 1489-1501.
- [34] XIAO J X, HAYS J, KRISTA A, et al. Sun database: large-scale scene recognition from abbey to zoo[C]// IEEE Conference on Computer Vision & Pattern Recognition, June 13-18, 2010, San Francisco, USA. Piscataway: IEEE Press, 2010: 3485-3492.
- [35] YANG Y Z, CHING L T, HAL D III, et al. Corpus-guided sentence generation of natural images[C]// Conference on Empirical Methods in Natural Language Processing, July 27-29, 2011, Edinburgh, UK. New York: ACM Press, 2011: 444-454.
- [36] ZHANG F, DU B, ZHANG L P. Saliency-guided unsupervised feature learning for scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4): 2175-2184.
- [37] ZHANG L, ZHEN X, SHAO L. Learning object-to-class kernels for scene classification[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2014, 23(8): 3241-3253.

[作者简介]



刘文华（1990—），女，北京交通大学博士生，主要研究方向为计算机视觉、大数据分析、数据挖掘。



李滢东（1982—），男，博士，北京交通大学计算机学院教授、副院长，主要研究方向为智能大数据分析、隐私保护、并行计算、计算机视觉。



王涛（1980—），男，博士，北京交通大学计算机与信息技术学院副教授，主要研究方向为机器学习与计算机视觉。



邬俊（1981—），男，博士，北京交通大学计算机与信息技术学院副教授，主要研究方向为机器学习理论及其在信息检索与推荐系统中的应用。



金一（1982—），女，博士，北京交通大学计算机与信息技术学院副教授，博士生导师，主要研究方向为以异质人脸识别为代表的计算机视觉、模式识别理论和方法。