基于多视图循环神经网络的三维物体识别

董 帅，李文生，张文强，邹 昆*

(电子科技大学中山学院 广东 中山 528406)

【摘要】对于三维物体的识别任务，基于多视图卷积神经网络的方法（MVCNN）在准确性和训练速度等方面都优于基于三维数据表示的方法。但 MVCNN 依赖于三维模型，且采用了固定视角的视图，不符合实际的应用场景；此外，其视图特征融合采用了最大值池化操作，会损失部分原始特征信息。针对这一问题，该文提出了一种基于多视图循环神经网络（MVRNN）的三维物体识别方法，从 3 个方面对 MVCNN 进行改进。首先，在交叉熵损失函数中引入特征辨识度指标，以提高不同物体特征之间的辨识度；其次，使用循环神经网络代替 MVCNN 的最大值池化操作来融合多个自由视觉视图特征，得到一个更加紧凑且物体外观信息完备的融合特征；最后，利用二分类网络对自由视角单视图特征和融合特征进行匹配，实现三维物体的细粒度识别。为了验证 MVRNN 的性能，分别在公开数据集 ModelNet 和自建数据集 MV3D 上进行对比实验。实验结果表明，与 MVCNN 相比，MVRNN 提取的多视图特征具有更高的辨识度，在两个数据集上的识别准确率均较有明显提升。

关 键 词 三维物体； 特征提取； 特征融合； 图像检索； 多视图

中图分类号 TP391.4 **文献标志码** A doi:10.12178/1001-0548.2019017

Recognition of 3D Object Based on Multi-View Recurrent Neural Networks

DONG Shuai, LI Wen-sheng, ZHANG Wen-qiang, and ZOU Kun*

(Zhongshan Institute, University of Electronic Science and Technology of China Zhongshan Guangdong 528406)

Abstract Multi-view convolutional neural networks (MVCNN) is more accurate and faster than those methods based on state-of-the-art 3D shape descriptors in 3D object recognition tasks. However, the input of MVCNN are views rendered from cameras at fixed positions, which is not the case of most applications. Furthermore, MVCNN uses max-pooling operation to fuse multi-view features and the information of original features may be lost. To address those two problems, a new recognition method of 3D objects based on multi-view recurrent neural networks (MVRNN) is proposed based on MVCNN with improvements on three aspects. First, a new item which is defined as the measure of discrimination is introduced into the cross-entropy loss function to enhance the discrimination of features from different objects. Second, a recurrent neural networks (RNN) is used to fuse multi-view features from free positions into a compact one, instead of the max-pooling operation in MVCNN. RNN can keep the completeness of information about appearance feature. At last, single view feature from free positon is matched with fused features via a bi-classification network to attain fine-grained recognition of 3D objects. Experiments are conducted on the open dataset ModelNet and the private dataset MV3D separately to validate the performance of MVRNN. The results show that MVRNN can exact multi-view features with higher degree of discrimination, and achieve higher accuracy than MVCNN on both datasets.

Key words 3D object; feature extraction; feature fusion; image retrieval; multi-view

近 5 年，基于深度学习的计算机视觉技术[1] 飞速发展，已广泛应用于智能安防和无人驾驶等多个领域。在大规模目标数据集中，针对具体的识别或检测任务，深层卷积网络可以通过端对端的方式自适应地学习如何从输入数据中提取和抽象特征，以及如何基于该特征进行决策。深层卷积网络既可作为图像特征提取和分类操作的统一体，又可以只作为特征提取网络供实例检索任务使用[2-3]。目前大多数基于深度学习的图像分类网络和目标检测框架都是针对二维图像提出的，但随着深度学习逐步应用到机器人导航和无人超市等领域，三维物体的识别技术也逐渐得到了研究人员的广泛关注。与二维图

收稿日期：2019－01－09；修回日期：2019－11－22
基金项目：国家青年科学基金 (61502088)；广东省自然科学基金 (2016A030313018)；广东省高等学校优秀青年教师培养计划 (Yq2013206)
作者简介：董帅 (1986－)，男，博士，主要从事机器学习、智能优化和先进控制等方面的研究.
通信作者：邹昆，E-mail：cszoukun@foxmail.com

像相比，三维物体识别的难点在于，同一物体的不同侧面可能存在较大差异，从不同角度观察会呈现出不同的形态，而不同物体在某个侧面上的差异可能很小，甚至呈现出相同的形态。这使得直接使用单视图 (即二维图像或投影) 分类网络的识别效果较差。

在深度学习受到广泛关注之前，有许多学者采用了 SURF 等传统几何方法[4-7] 对三维物体的识别技术进行了探索，取得了一定的成果，但这类方法的鲁棒性和泛化能力较差。近几年，研究者逐渐将深度学习推广到三维物体识别领域，提出了多种方法。这些方法可以大致分为两类：基于三维数据表示的方法和基于多视图表示的方法。文献 [8] 提出了基于体素网格和三维卷积的 VoxelNet，该网络是二维平面卷积到三维空间卷积的直接推广，由于计算量过大，输入模型的体素分辨率一般较低，进而导致识别精度也较低。文献 [9-10] 提出了针对三维点云的 PointNet 及后续的一系列方法，这些方法基于点云的无序性提出多种非欧卷积网络[11-12]，具有较大的影响力，但同样存在计算量大和训练困难的问题。文献 [13] 提出了基于 SSD 的 6 维位姿估计目标检测框架，开创性地将位姿估计和目标检测二者结合，具有启发性。文献 [14] 提出了基于深度霍夫投票的 3D 目标检测框架 VoteNet，该框架主要用于场景的识别，未关注单个实例的分类和检索问题。文献 [15] 提出的基于多视图的卷积神经网络 (MVCNN)，与基于三维数据的方法并行。MVCNN 在分类和检索任务上的表现均优于基于三维数据的识别方法。在文献 [16] 中，对 MVCNN、PointNet++和 VoxelNet 等多种方法进行对比，并指出多视图方法的优异表现主要得益于庞大的二维图像数据集。但 MVCNN 存在两个方面的不足：1) 依赖于精确的 3D 模型，且采用了固定视角的视图，这并不符合真实的应用场景，导致算法泛化能力不足；2) 采用了最大值池化操作来对多视图进行融合，融合后的特征会损失大量信息。

针对 MVCNN 存在的问题，本文提出了一种基于 MVRNN 的三维物体识别方法。首先，设计了一个包含特征辨识度指标的目标函数用于训练网络，能够得到辨识度更高的物体单视图特征和融合特征；其次，使用循环神经网络 (recurrent neural network, RNN) 对多个视图特征进行融合，得到一个更加紧凑且包含更丰富信息的融合特征作为物体的注册特征；最后，利用单视图特征对注册特征进行检索。与 MVCNN 相比，MVRNN 存在以下优点：1) 不依赖于 3D 模型，在实际应用中，只需要采集 2D 图片提取特征并进行融合；2) 对视图的视角和数量没有要求，对不同视图的特征信息利用更充分；3) 利用循环结构网络进行特征融合，兼具紧凑性和完备性。

1 问题描述

1.1 多视图数据集

PASCAL3D+和 Tsukuba 等公开三维数据集主要针对三维模型的分类，并不适用于多视图的识别场景。文献 [15] 基于 ModelNet 建立了多视图的数据集，但只采用了图 1 所示的 12 个固定位置和视角，并不完全符合实际应用的场景。为了充分展现 MVRNN 的优点，本文自建数据集 MV3D (multi-view 3D) 用于对比试验。

图 1 ModelNet 数据集固定视角示例

MV3D 采用 Unity 软件制作，将三维模导入软件，并在 Camera 的视场中随机平移和旋转模型，得到二维视图。该数据集共有 95 个三维物体模型，每个物体包括 100 个二维视图。物体模型较 ModelNet 更加精致，纹理也更加丰富。该数据集中存在一些在不同视角角度下外观差异较大的物体，以及一些属于不同类别但在某些视角下形态十分相近的物体。图 2 展示了该数据集中的部分样本。

图 2 MV3D 数据集示例

1.2 MVRNN 三维物体识别框架

记三维物体的集合$O=\{o_i,i=1,2,\cdots,l\}$，o_i对应的注册视图集合为$V_i=\left\{v_j^i,j=1,2,\cdots,m_i\right\}$，检索视图集合为$U_i=\left\{u_k^i,k=1,2,\cdots,n_i\right\}$。构建如图 3 所示的三维目标识别框架，该框架包含特征提取模块

$E(\cdot)$、特征融合模块$F(\cdot)$、分类输出模块$C_1(\cdot)$和$C_2(\cdot)$、检索匹配模块$M(\cdot)$5 个网络模块。

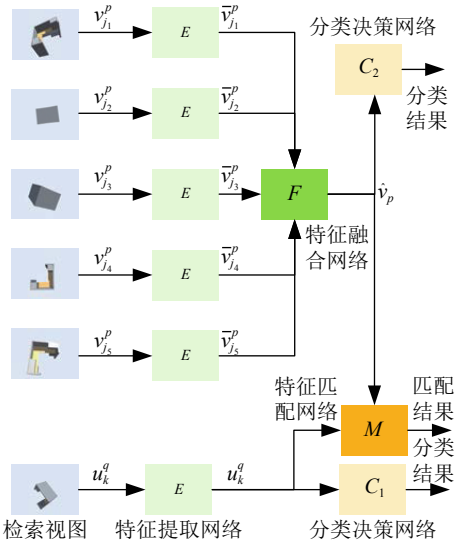

图 3　三维物体识别框架

特征提取模块$E(\cdot)$的输入为单注册视图v_j^i或单检索视图u_k^i，输出为对应的单视图特征$\bar{v}_j^i = E\left(v_j^i\right)$或$\bar{u}_k^i = E\left(u_k^i\right)$；特征融合模块$F(\cdot)$的输入为同一物体的多个单视图特征$\left(\bar{v}_{j_1}^i, \bar{v}_{j_2}^i, \cdots, \bar{v}_{j_m}^i\right)$，输出为新的注册特征$\hat{v}_i = F\left(\bar{v}_{j_1}^i, \bar{v}_{j_2}^i, \cdots, \bar{v}_{j_m}^i\right)$；$C_1\left(\bar{v}_k^p\right)$和$C_2\left(\hat{v}_p\right)$分别为单视图特征和多视图融合特征的分类网络，用于预测视图所属物体类别；匹配模块$M(\cdot)$是一个二分类网络，输入为注册特征$\hat{v}_p$和检索特征$\bar{u}_k^q$，用于预测二者是否属于同一物体。

在 MVCNN 中，$F(\cdot)$采用了简单的最大值池化；此外，$F(\cdot)$还可采用均值池化和直接拼接等实现方法。本文利用 RNN 代替最大值池化实现特征融合，此即为 MVRNN 的由来。

由于多个模块同时训练难度较大，整个框架采用分步训练的策略：1) 训练分类分支$E(\cdot)$和$C_1(\cdot)$，固化$E(\cdot)$并提取单视图特征；2) 训练分类分支$F(\cdot)$和$C_2(\cdot)$，固化$F(\cdot)$计算融合特征；3) 训练二分类网络$M(\cdot)$。$C_1(\cdot)$和$C_2(\cdot)$只用于$E(\cdot)$和$F(\cdot)$的训练，并不直接参与预测。

2　MVRNN 具体实现方案

2.1　特征提取网络

与 MVCNN 一样，在 MVRNN 中$E(\cdot)$和$C_1(\cdot)$直接采用了 ResNet-18[17] 的结构，并加载了预训练的参数进行微调。输入图片尺寸为 224*224，输出特征长度为 512。训练时，采用的损失函数为：

$$L_{\text{total}} = L_{\text{cls}}(p,q) + \mu\|w\|^2 + \lambda L_{\text{rect}}(p,q) \tag{1}$$

式中，p为视图的真实类别；q为预测类别；$L_{\text{cls}}(p,q) = \frac{1}{m}\sum_{i=1}^{m}(-p^i \log q^i)$为交叉熵损失函数；$m$ 为每一批次的样本数量；$\mu\|w\|^2$为 L2 正则项；$\lambda L_{\text{rect}}(p,q)$为矫正项；$\lambda$和$\mu$为可调权重，本文分别取 1 和0.000 1。$L_{\text{rect}}$的定义为：

$$L_{\text{rect}}(p^i,q^i) = \frac{1}{m}\sum_{i=1}^{m}\max(0,\text{rect}(p^i,q^i))$$

$$\text{rect}(p^i,q^i) = \begin{cases} \log\left(\dfrac{1+s_{1,q=p}^i+\xi}{s_{2,q=p}^i+\xi}\right) & p^i = q^i \\ 0 & p^i \neq q^i \end{cases} \tag{2}$$

式中，$s_{1,q=p}^i$表示第i个样本为真实类别的概率；$s_{2,q=p}^i$表示第i个样本的次大预测概率；ξ取 0.01。L_{rect}会在输入的视图被正确分类的情况下进一步微调模型参数，使得真实类别的预测概率更接近 1，同时其他类别的概率更接近 0，进而增大所提取到特征的辨识度。

2.2　循环多视图特征融合网络

特征融合网络的作用是对多个视图特征进行融合，得到一个能够完整描述物体形状和纹理信息的特征。本节借鉴视频分析方法，采用 RNN 来融合特征，其结构如图 4 所示。物体的多个视图在时间上无相关性，但在空间上是关联的，因此能够借助 RNN 的记忆能力来融合特征。

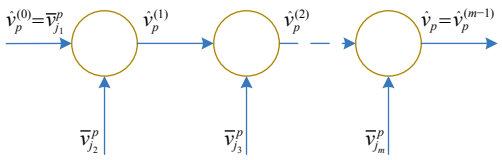

图 4　多视图特征融合网络

$F(\cdot)$网络采用图 5 所示的结构，每个循环体中包括线性全连接层 (full connection, FC) 和双曲正切单元 Tanh，最后的分类层$C_2(\cdot)$包括了线性全连接 FC 和 Softmax 操作，全连接层神经元数量均为 1 024，融合后特征长度为 512。$F(\cdot)$循环体的数量可以随输入视图的数量变化，即输入视图数量不固定。$F(\cdot)$的训练同样采用了式 (1) 所示的损失函数，λ取 0.01，μ取 0。

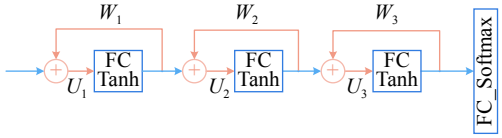

图 5　聚合网络模型

2.3　检索匹配网络

相似度匹配模块$M(\cdot)$是一个二分类模型，使用

了三层的全连接神经网络结构，输入由单视图特征和融合后特征拼接而成，隐藏层由线性全连接、Batch Normalization 和 ReLU 组成，输出层由线性全连接 FC 和 Softmax 组成，隐藏层神经元数量均为 1 024，网络结构如图 6 所示。$F(\cdot)$的训练同样采用了式 (1) 所示的损失函数，其中，λ取 0.000 5，μ取 0。

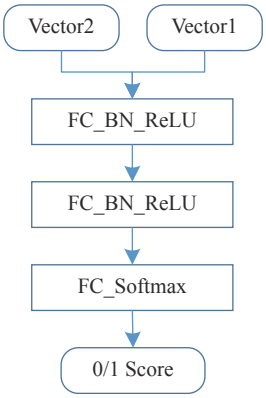

图 6　相似度匹配模型

3　特征融合方法对比

特征融合是传统机器学习中比较常用的手段，一般需要根据先验知识来提取不同类别的特征信息，并进行协同决策。特征融合在深度学习领域也得到广泛应用，比如 ResNet 的残差模块和 DenseNet 的跨层连接，都对不同层的特征进行了融合。常见的特征融合方法包括直接拼接 (concatenating)、堆叠 (stacking)、相加 (adding)、最大值池化 (max-pooling) 和均值池化 (average-pooling) 等。其中，堆叠可以看做是直接拼接的特例，相加则等效于均值池化。衡量特征融合方法的主要准则有两个：1) 原始特征的信息是否会损失，即信息的完备性；2) 融合后特征是否便于后续计算，即特征的紧凑性，一般指融合特征的长度。此外，传统机器学习的特征融合还比较注重被融合特征之间的差异性，差异越大则信息量越多，但该准则对于本文所解决的问题并不适用。

对于三维物体的多视图特征融合任务而言，直接拼接能够保证信息的完备性，但融合后特征长度较大，会导致网络规模较大，且训练难度增大；最大值池化和均值池化得到的特征比较紧凑，但会损失部分信息；而 RNN 则兼具完备性和紧凑性。几种方法得到的融合特征长度比较直观，直接拼接方法的完备性也是毋庸讳言。

为了对比两种池化方法和 RNN 的完备性，本节设计了一个比较极端的二维特征融合任务，对比结果如图 7～图 10 所示。图 7 包含 10 个物体的不同视图特征，每条曲线表示一个物体，曲线上的点表示不同视图的特征。特征空间可以分为左上、左下、右上和右下 4 个子空间，子空间内的物体特征存在较大的相似性。从每条曲线随机抽取 5 个点进行融合，重复得到融合特征的分布。最大值池化的结果如图 8 所示，其中，左下两个物体特征出现了重叠，右上的类似。均值池化的结果则是左上和右下的物体特征出现重叠，具体如图 9 所示。RNN 采用了单隐含层 10 神经元的全连接网络，其融合结果如图 10 所示。RNN 引入了新的网络层将特征映射至新的空间，10 个物体被有效区分。

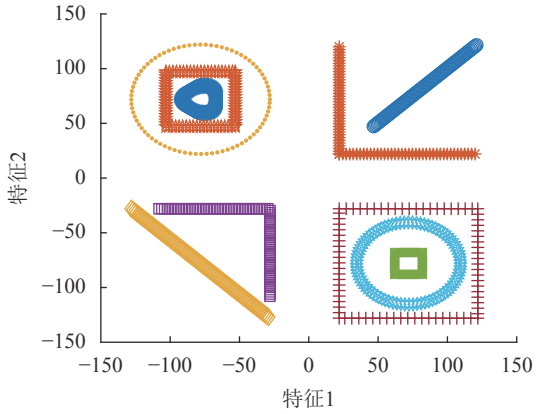

图 7　原始特征分布

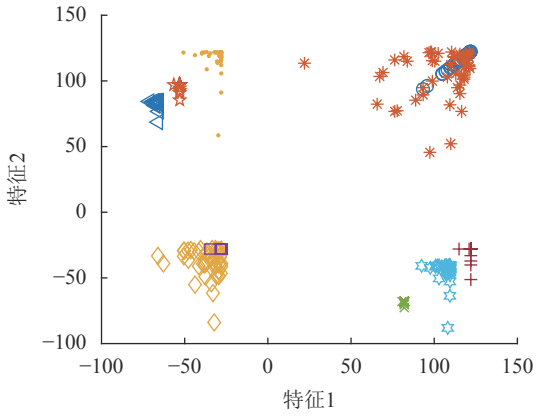

图 8　采用最大值池化进行融合后的特征分布

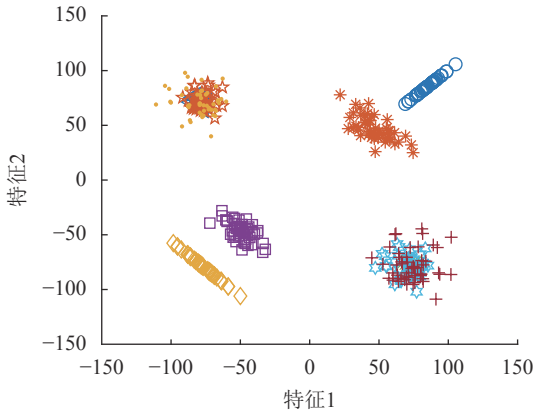

图 9　采用均值池化进行融合后的特征分布

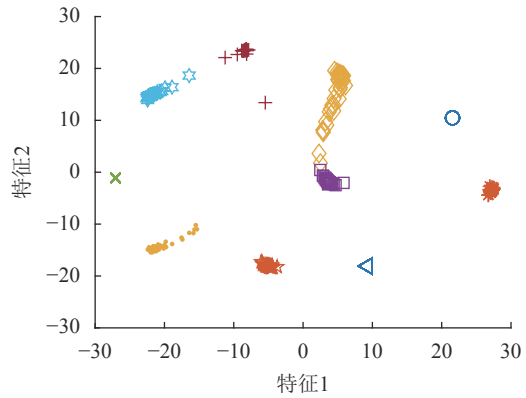

图 10　采用 RNN 进行融合后的特征分布

4　实验结果与分析

为了说明 MVRNN 在融合多视图特征上的优越性，本节在 ModelNet 数据集[15]和自建数据集 MV3D 上进行了多组对比分析。

从 ModelNet 数据集随机抽取 4 000 个物体，每个物体分别抽取 6 张和 12 张视图，按照 6:1:3 的比例划分训练集、验证集和测试集。MVCNN 和 MVRNN 在融合特征分类任务和实例检索任务上的性能如表 1 所示。从表 1 可以看出，相较于 MVCNN，MVRNN 在分类任务上有一定的提升，且融合的视图越多，二者的准确率都有提升；在检索任务上，MVRNN 明显优于 MVCNN；需要注意的是，随着视图的增多，MVCNN 检索的准确率会下降，这是由于 ModelNet 数据集中模型本身都比较简单粗糙，缺乏具有辨识度的纹理，最大值池化操作更容易丢失信息，图片越多，更有可能导致部分关键信息的丢失。

表 1　MVCNN 和 MVRNN 准确率对比 (ModelNet)

任务描述	准确率/%	
	MVCNN	MVRNN
6视图分类任务	93.2	94.3
6视图检索任务	96.1	99.0
12视图分类任务	95.4	95.9
12视图检索任务	94.3	100.0

由于 ModelNet 数据集具有固定视角的限制，无法充分验证 MVRNN 的性能。因此，本文利用 Unity 3D 制作了 MV3D 数据集，其中训练集包含 65 个物体，测试集包含 30 个物体，每个物体包含 100 张视图。数据集的设定如下：

1) 训练特征提取网络时，训练集中所有的视图 (6 500 张) 全部参与训练。

2) 训练特征融合网络时，从每个物体随机抽取 6 个单视图特征构建六元组作为网络输入；训练集包含 65 个物体，每个物体包含 2 000 个六元组；测试集由同样的 65 个物体生成，每个物体包含 500 个六元组。即训练样本数量为 130 000，测试样本数量为 32 500。

3) 训练匹配网络时，从物体 A 随机抽取 7 个单视图特征 A1～A7，从物体 B 抽取 1 个单视图特征 B1，构建正负两个七元组样本作为网络输入，其中 A1～A6 输入特征融合网络生成融合特征，A7 为检索特征正样本，B1 为检索特征负样本；训练集包含 65 个物体，每个物体包含 2 000 个七元组；测试集包含 30 个物体，每个物体包含 2 000 个七元组。即训练样本数量为 130 000，测试样本数量为 60 000。

在 MV3D 数据集上进行 7 种方法的对比测试，结果如表 2 所示。实验的设定如下：1) 基于单个视图特征$\tilde{v}_{ji}^{p}$进行分类和检索；2) 基于多个单视图特征$\tilde{v}_{ji}^{p}$进行单独匹配，并取置信度最高的视图作为最终匹配结果；3) MVCNN，即$F(\cdot)$为最大值池化；4) $F(\cdot)$为直接拼接；5) $F(\cdot)$为均值池化；6) MVRNN without L_{rect}；7) MVRNN with L_{rect}。各组实验涉及到的卷积网络和相似度匹配模型均采用同样的结构，且所有模型均使用相同的训练方法和超参，batch_size 为 50，采用 Nesterov[18] 梯度加速算法，初始学习率为 10^{-2}，稳定后变为 10^{-3} 和 10^{-4}，动量为 0.9，dropout 概率[19] 为 0.3。top1_dst定义为在检索正确的结果中，1.0 与最大相似度之间的平均距离，即$\frac{1}{k}\sum_{i\in\text{correct_retrival}}(1.0-s_{1,q=p}^{i})$；top2_dst则表示在检索正确的结果中，最大与次大相似度之间的平均距离，即$\frac{1}{k}\sum_{i\in\text{correct_retrival}}(s_{1,q=p}^{i}-s_{2,q=p}^{i})$；$\log\left(\frac{\text{top2_dst}}{\text{top1_dst}}\right)$可以衡量特征辨识度的高低，top2_dst越大，同时top1_dst越小，则该值越大，也说明特征的辨识度越高。

表 2　MVRNN 性能对比 (MV3D)

实验	acc/%		特征辨识度		
	分类	检索	top2_dst	top1_dst	$\log\frac{\text{top2\_dst}}{\text{top1\_dst}}$
1	80.737	64.197	0.149 76	0.014 04	1.028 03
2	80.737	72.737	0.041 19	0.003 32	1.093 65
3	100.000	80.452	0.328 26	0.015 80	1.317 56
4	98.175	83.530	0.223 67	0.008 82	1.404 14
5	99.965	85.254	0.318 46	0.007 64	1.619 96
6	100.000	86.540	0.216 93	0.002 00	2.035 29
7	100.000	89.080	0.234 95	0.001 75	2.127 94

从表 2 来看，MVRNN 准确率最高，即使损失函数不考虑L_{rect}项，结果依然较其他方法好。最大值池化、均值池化和直接拼接 3 种方式准确率相近，为第 2 梯队；只使用单视图的两种方法效果最差。

在目标函数中增加L_{rect}项后，MVRNN 在单视图分类和融合特征检索的准确率上都得到了明显提升，具体结果如表 3 所示。结合表 2 的特征辨识度指标来看，L_{rect}能够提升特征辨识度，进而提升分类和检索的准确率。

表 3　L_{rect} 效果对比

实验	acc/%		
	单视图分类	融合特征分类	融合特征检索
6	80.737	100.000	86.540
7	82.281	100.000	89.080

为了进一步对比 MVRNN 和 MVCNN 的性能，本节对表 2 中的实验 3 和实验 7 进行扩展，得到了视图数量分别为 2，4，6，8，10 时，训练集物体数量为 10，20，30，40，50，65 时的检索准确率，具体结果如图 11 所示。从图中可以看出：1) 随着训练集物体数量的增加，检索准确率也不断增加；2) 在物体数量超过 30 后，准确率整体的提升幅度较小，物体数量为 30 时对应的训练样本数量为 60 000；3)MVRNN 整体准确率较 MVCNN 高约 8%。

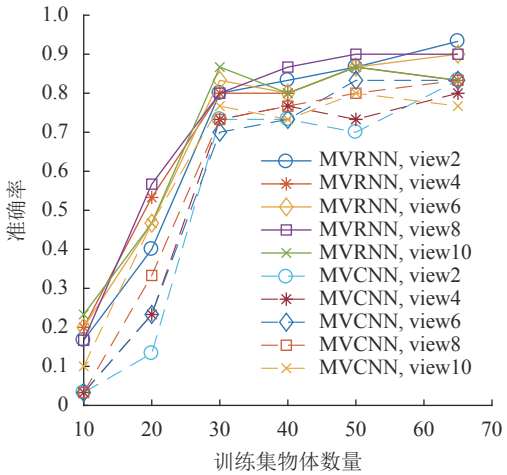

图 11　训练样本数量对训练结果的影响

本文还基于 MVRNN 开发了一个简单的商品识别系统。该系统包括商品注册和商品识别两个模式，具体应用如图 12 和图 13 所示。在注册阶段，采集商品实例的不同视图，以提取视图特征进行融合，并对融合特征进行注册；在识别阶段，则根据商品单视图特征对融合特征进行检索。在 100 种常见饮料食品类商品上进行测试，注册图片不超过 9 张，即可完成大部分商品实例的检索，准确率约为 90%。

图 12　商品注册

图 13　商品识别

5　结 束 语

针对三维物体的分类和检索问题，本文对 MVCNN 进行改进，提出了 MVRNN。通过在损失函数中引入特征辨识度指标，能够有效提升分类和检索的准确率；利用 RNN 代替最大值池化操作，使得融合特征具有信息完备性。在 ModelNet 数据集和 MV3D 数据集上，MVRNN 的表现较 MVCNN 有了明显提升。在未来的研究中，拟制作大规模商品数据集以开展 MVRNN 的应用研究；此外，将 MVRNN 与 SSD 等目标检测框架结合来估计物体的六维位姿也是一个比较有前景的方向。

参 考 文 献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] WAN J, WANG D, HOI S C H, et al. Deep learning for content-based image retrieval: A comprehensive study[C]//The 22nd ACM International Conference on

Multimedia. New York: ACM, 2014: 157-166.
[3] YAO H, ZHANG S, ZHANG Y, et al. One-shot fine-grained instance retrieval[C]//The 25th ACM International Conference on Multimedia. New York: ACM, 2017: 342-350.
[4] KAZHDAN M M, FUNKHOUSER T A, RUSINKIEWICZ S. Rotation invariant spherical harmonic representation of 3D shape descriptors[C]//The 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Goslar: Eurographics Association, 2003: 156-164.
[5] KNOPP J, PRASAD M, WILLEMS G, et al. Hough transform and 3D SURF for robust three dimensional classification[C]//The 2010 European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2010: 589-602.
[6] CHAUDHURI S, KOLTUN V. Data-driven suggestions for creativity support in 3D modeling[J]. ACM Transactions on Graphics, 2010, 29(6): 183.
[7] KOKKINOS I, BRONSTEIN M M, LITMAN R, et al. Intrinsic shape context descriptors for deformable shapes[C]//The 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 159-166.
[8] WU Z, SONG S, KHOSLA A, et al. 3D ShapeNets: A deep representation for volumetric shapes[C]//The 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1912-1920.
[9] QI C R, SU H, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//The 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 77-85.
[10] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]//The 2017 Neural Information Processing Systems Conference. Nice: Curran Associates. 2017: 5099-5108.
[11] WANG Y, SUN Y, LIU Z, et al. Dynamic graph CNN for learning on point clouds[J]. ACM Transactions on Graphics, 2019, 38(5): 146.
[12] HUA B, TRAN M, YEUNG S. Pointwise convolutional neural networks[C]//The 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 984-993.
[13] KEHL W, MANHARDT F. SSD-6D: Making RGB-Based 3D detection and 6D Pose estimation great again[C]//The 2017 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 1530-1538.
[14] QI C R, LITANY O, HE K, et al. Deep hough voting for 3D object detection in point clouds[C]//The 2019 International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 9277-9286.
[15] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition[C]//The 2015 International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 945-953.
[16] SU J, GADELHA M, WANG R, et al. A deeper look at 3D shape classifiers[C]//The 2018 European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2018: 545-561.
[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//The 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.
[18] SUTSKEVER I, MARTENS J, DAHL G E, et al. On the importance of initialization and momentum in deep learning[C]//The 30th International Conference on Machine Learning. Cambridge, MA: MIT Press, 2013: 1139-1147.
[19] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

编　辑　叶　芳