指导教师：杨涛　　提交时间：2016/3/14

CVPR2015 Paper Translation

No: 01

姓名：谭冲

学号：2013302632

班号：10011306

学习通过移动

摘要

当前的主导范式特征学习在计算机视觉领域依赖使用数以百万计的手标记图像训练神经网络，来完成目标识别的任务。使用其他任何形式的监督来学习功能多样化的视觉任务是有可能的吗?在生物学中,生物发展视觉感知能力的目的是在世界上移动和表演。从这个观察的灵感出发,在此项工作中我们调查了运动意识(即自我运动)是否可以用作一个对于特征学习的监督信号。与关于运动的分类知识和信息是免费的移动代理这一观点是截然相反的。我们发现在使用相同数量图像的训练下,使用运动作为监督的特征学习比使用分类标号作为监督的特征学习在场景识别、物体识别、视觉测程和关键点匹配任务上更有优势。

“我们必须为了移动而感知,但我们也必须为了感知而移动.”

--J.J Gibson

1.简介

最近计算机视觉的发展表明,为了完成目标识别任务，通过神经网络训练的可视化特征学习在使用了超过一百万标签图像后，对大量计算机视觉任务，如语义分割，,对象检测和行动分类是有用的。然而,目标识别是许多视觉任务中的一个。例如,人类用视觉感知识别对象,理解场景的空间布局和执行操作,如在地上移动。目标识别有什么特别之处或者它是可以通过其他监督方式的有用的视觉表征吗?显然,生物执行复杂的视觉任务,他们不太可能需要用数以百万计标签形式的外部监督。未标记的视觉数据是容易获取到的,在理论上这些数据可以被用来学习有用的视觉表征。然而,到目前为止无监督的学习方法尚未实现,没有应用程序付诸于复杂而又真实的实际图像。

生物使用感知系统获取在他们的周围让他们采取行动,实现他们的目标的感官信息。生物和机器人使用他们的运动系统对他们的环境作出反映。这些物体使用他们自己的运动系统作为一种监督方式来获取有用的知

觉表征是有可能的吗?运动知觉理论有着悠久的历史,但是没有利用运动信息系统地阐述感知计算模型。在工作中，我们关注视觉感知和为了获取有用的视觉信息基于运动（即自我运动)提出一个模型。当我们提及有用的视觉表征[34],我们的意思是它应具有以下两个特点——(1)执行多个视觉任务的能力,(2)通过外界指引者提供的仅有几个标志执行新视觉任务的能力。

移动机器人通过自己的运动系统，自然地意识到自己的运动(即自动)。换句话说,运动是自然而然的。例如,许多哺乳动物的前庭神经系统为它们提供了方向感。在人类和其他动物中,大脑获取关于眼球运动和动作执行的信息[9]。移动机器人可以估计其行为，从电动机发出命令或从安装在机器人上的测距传感器如陀螺仪和加速度计，使它移动。

我们认为通过执行简单的与运动相关的任务能获得有用的视觉特征。移动机器人是可以看做移动的相机,因此世界上运动的知识类似于相机运动的知识。使用这一观点,我们提出了关于机器人在移动的时候收到连续的图像的视觉转换问题。直观地,预测在两个图像之间的相机转换任务应该强制代理机器去学习熟悉的存在于两个图像之间的识别视觉元素的特征(即视觉通信)。在过去,如手动的寻找一致性的尺度不变特征转换等任务的特性被发现对于对象识别也是非常有用的，这表明基于学习的运动也产生了对于这个任务很有用的特性。

为了验证我们关于利用运动的特征学习的假设，,我们训练了多层神经网络去预测相机对图像之间的转换。为了观念的论证,我们首先在MNIST数据库中证明了我们方法的有效性。我们展现了当可用的类标记的数量有限时，使用我们方法的特征学习比使用之前无监督的特征学习更好接下来,我们将在实际的图像世界中评估我们的方法。为了达到这个目的,我们使用了来自 KITTI 和旧金山数据库的部分图像及一辆穿越城市的汽车数据记录仪。这些数据模拟了一个代理机器在世界上移动场景。根据这些数据产生的特性学习被用来评估四个任务(1)在太阳下的场景识别,(2)视觉测距(3)关键点匹配和(4)图像中的对象识别。

我们的结果表明,在相同数量的训练数据下使用运动作为监督的特征学习比使用类标签作为监督的特征学习更加有利。我们还表明,基于训练的运动方法优于以前的无视频图像监督缓慢的特征分析的方法。对于我们最好的知识而言,这项工作第一次提供了有效的证明从现实世界非视觉的运动信息中能获得视觉表征。

本文的其余部分组织如下:在第二部分我们讨论相关工作；在第三,四、五部分，我们展示目前的方法,数据库的细节，以及我们将在第六部分的作出的结论与讨论。

2. 相关工作

过去的无监督学习已经成为学习姿态特征的发现足以重建图像的紧凑而丰富的图像表征问题的主流方法。另一条工作线已经集中在从视频或从图像中不变的转换的特征学习。[通过使用玻耳兹曼机的模拟空间改建获取的特性学习,不要评估其质量。

尽管有很多无监督学习的工作,但这种方法要适用于复杂现实世界的图象是有待开发。无监督学习的替代品是使用可以自由获取的内部有用信号的特征学习。例如,使用内在有用的信号可用于机器人预测可遍历路径的特性学习，这些机器人是受过神经网络训练，可直接通过视觉输入驱动行为的

在这个工作中,我们建议使用非视觉的运动信息作为视觉特征学习自我监督的一种形式。不像其他以前的工作,我们将展示我们的方法适用于真实世界的图像。最接近我们的方法是建立使用运动重建来自输入源图像转换后的图像的自动转换机。这项工作是在自然界中纯粹的概念,没有特征学习的质量评估。相比之下,我们的方法是通过使用暹罗像网络模型来预测两个图像之间的转换的运动监督。

我们的方法也可以被视为一个实例的视频特性学习。从视频中执行特征学习通过强加的约束，时间接近帧应该有类似的特征表示（即缓慢的特征分析），而不占任何场景中的摄像机运动或物体的运动。我们的主要观察是摄像机的运动（即运动）对于移动机器人是容易获得的，可作为一个强大的自我监督的来源。

3.一个基于运动学习的简单模型

我们用一个卷积神经网络模型模拟机器人的的视觉系统。机器人优化它的可视化表征通过在从其电机系统中获得的运动信息（即相机转换）和只使用它的视觉输入的行为预测之间的最小化误差。执行这项任务相当于训练一个 CNN 的双流，以两个图像作为输入，预测运动，代理机器的行为是在两幅图像中获取的位置之间移动。为了学习有用的视觉表征，机器人连续地在它的环境中执行此任务。

在这项工作中，为了评估特征学习的实用性，我们使用微调的训练程序范例。训练是一个随机初始化 CNN 的权值，达到最优的过程，是不一样的目标的辅助任务。

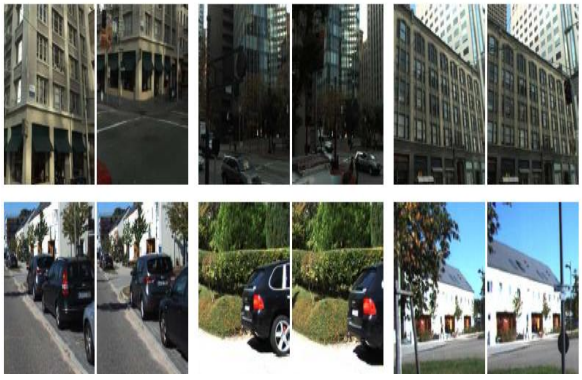


图 1：为获得有用的视觉特征，探索运动监督的实用性。一个装配了视觉传感器的移动机器，当他在其环境中移动时接收到一系列的图像序列作为输入。这个机器的移动相当于一个摄像机的运动。在这项工作中，提出了基于学习的运动作为预测图像对的变换的问题。图中的顶部和底部的显示一些来自于 SF 和 KITTI 数据库的样本图像对，这些图像对被用于特征学习任务。微调是修改一个用于给定对象训练的 CNN 权值的一个过程。我们的实验是比较对于多目标任务而言，使用基于运动训练的特征学习和基于标签的特征学习及缓慢特征训练的学习的实用性。

3.1 两个流体系结构

CNN 的每一个流独立计算每一个图像的特征。这两个数据流共享相同的架构和相同的权重，从而执行相同的操作，用于计算特征。个别的数据流被称为基本-CNN。将两个 BCNN 得到的特征级联并通过下游进入另一个称为顶级的 CNN（见图 2）。TCNN 是负责利用的 BCNN 特征去预测在输入图像对之间的图像转换。经过训练后，TCNN 被移除，单一的 bcnn 是作为一个标准的 CNN，用于目标任务的特征计算。

3.2 简称 CNN 架构

缩略语 Ck，FK，P，D，OP 是一个 k 过滤器的卷积层，一个分别与 k 过滤器合并，差和输出（运算）完全连接的层。我们用热鲁非线性运算作用于除了输出层外的每个卷积/完全连接层。移除层总是使用 0.5 的移除。输出层是一个完全连接的层，数据单位等于期望输出的数值。作为一个我们的符号例子，c96-p-f500-d 指网络 96 滤波器在卷积层中通过热鲁非线性，集中层，完全连接层伴随着 L1 转换 L2 Lk F1 F2 Base-CNN Stream-1 Base-CNN Stream-2 Top-CNN。

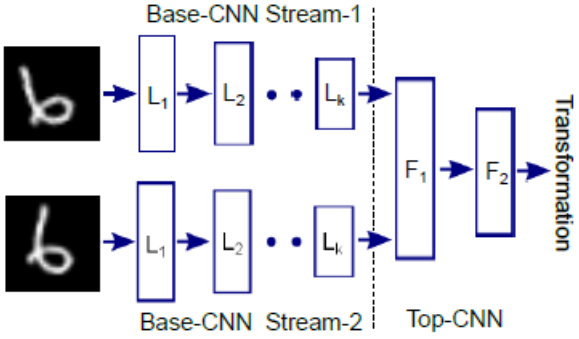


图 2：特征学习方法的描述。通过以两张图像作为输入的连体式卷积学神经网络能获得视觉特征，和预测图像之间的转换（即运动）。SCNN 的每个流（称为为基本—CNN 或 BCNN）计算一个图像的特征。两个 BCNN 的输出是被作为称为顶部—CNN（TCNN）（如层 F1，F2）的第二个多层 CNN 的输入而连接，通过。两个 BCNNS 具有相同的架构和共享权重。在特征学习后 TCNN 被丢弃和单个 bcnn 流被用作 CNN 来提取特征用于执行如场景识别的目标任务。500 单位,热鲁非线性和一个移除层。我们使用[17]训练我们所有的模型。

3.3 缓慢特性分析(SFA)基线

缓慢特征分析（SFA）是特征学习的一种及时基于有用特征缓慢变化的

方法，我们使用了以下 SFA 的公式来对比损失

$$L(x_{t_1}, x_{t_2}, W) = \begin{cases} D(x_{t_1}, x_{t_2}) & \text{if } |t_1 - t_2| \le T \\ 1 - max(0, m - D(x_{t_1}, x_{t_2})) & \text{if } |t_1 - t_2| > T \end{cases} \quad (1)$$

其中，L 是损失，XT1，XT2 是指在时间 T1、T2 分别观察帧的特征表示，W 是指定特征提取过程的参数，D 是衡量距离的参数，m 是一个预定义的属性和 T 是一个决定两帧是否临时关闭的预先确定的时间阈值。在这工作中，XT 是带有 W 和 D 权值的 L2 距离的 CNN 特征权重计算。SFA 训练使用双流架构，采取了图像对作为输入和作为两个流的对应输出的输出产生对 XT1，XT2。

3.4 使用 MNIST 的概念证明

在 MNIST 中，运动是由数字图像的随机变换综合构成的数据产生。在 60K 图像的训练集中，数剧是随机采样并转化的，使用两种不同数据集的随机变换生成的图像对。CNN 被训练用于预测这些图像对之间的转换。

3.41 MNIST 数据

对于基于运动的训练，数据之间的相互转化被限制在范围为[- 3，3]的一个整型值和相对旋转被限制在范围为[-30，30]度。转换预测被作为一个分类任务，有着三个独立的柔和的最大损失（每一个都为翻译，沿×，Y 轴和 Z 轴的旋转）。SCNN 被训练以尽可能减少这三种损失的总和。沿 X，Y 轴的转换被分别放进七个均匀的分割空间中。旋转的分级为 3 度一个空间，最后被放进了 20 空间中（或类）。对于 SFA 的基础训练，在区间[-1，1]转化和在区间[-3，3]度的相对转动的图像对，被认为在时间上彼此接近（见方程 1）。一共有 500 万个图像对被用来训练程序。

3.42 MNIST 网络架构

我们尝试了多种 BCNN 架构，分别为每个训练方法选择了最佳结构。对于基于运动的训练，两个 BCNN 流被 TCNN:f1000-d-op 连接。训练被 40K 次迭代执行（即 5M 的例子）使用被每 10000 迭代后的两个因素减少的 0.01 的初始学习率。

以下结构使用 bcnn-f500-d-op 进行了微调。为了评价 BCNN 特征的质量，将 bcnn 所有图层的学习率在数字分类的微整期间设置为 0。微调是以恒定值 0.01 的学习率进行 4K 次的迭代执行（相当于 10K 的标记训练的 50 次训练）。

3.43 MNIST 结果

bcnn 的特点是计算在数字分类的任务，使用 100，300，1K、10K 级类标记训练的例子中的误差评估。这些数据集是从 60k 标准训练的数据集中随机采样数据构造的。这部分实验，原始的数字图像被使用（即没有任何

转换或数据增强）。标准 10K 的数字测试集被用来进行评价和在表 1 中被报告的 3 次运行的平均错误率。

bcnn 的架构：c96-p-c256-p，被发现是最佳的基础运动和 SFA 训练也是从零开始（即随机重量初始化）的最佳训练。其他体系结构的结果在补充材料中提供。对于 SPF 基础的训练，我们尝试了多个 m 设定值，发现了在 m 等于 10，100 时产生了最好的结果。我们的方法优于卷积深度信念网络[22]，一个以前的基于学习特征不变量的转化方法[29]和基于 SFA 的训练。

4. 在自然环境中根据运动学习视觉特征

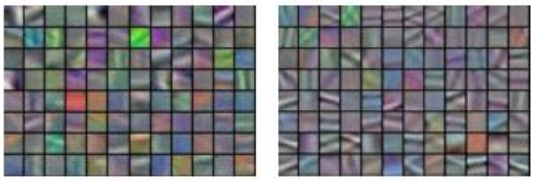
为了特征学习我们用两个主要的现实世界的数据来源：KITTI 和 SF 数据库，通过使用相机和安装在驾驶通过城市场景的汽车上的里程计传感器收集的数据。有关数据的详细信息，实验程序，网络体系结构及结果分别在 4.1、4.2、4.3 和 5 中提供。

4.1 KITTI 数据库

KITTI 数据库提供的测距和图像数据记录在由通过城市景观的汽车产生的 11 个可变长度的短行程中。在整个数据集的总帧数为 23201。

Method	# examples for finetuning			
	100	**300**	**1000**	**10000**
Autoencoder [17]	24.1	12.2	7.7	4.8
Ranzato et al. [29]	-	7.18	3.21	**0.85**
Lee et al. [22]	-	-	2.62	-
Train from Scratch	20.1	8.3	4.5	1.6
SFA (m=10)	11.2	6.4	3.5	2.1
SFA (m=100)	11.9	6.4	4.8	4.7
Egomotion (ours)	**8.7**	**3.6**	**2.0**	0.9

表格 1：在 MNIST 中，不同训练方法的比较显示有监督的运动训练比以前许多的无监督学习方法性能更优性能被报告为错误率。



(a) KITTI-Net (b) SF-Net

图 3：1 层可视化过滤器通过基于在 KITTI and SF 数据库训练的运动来学习。大多数的 1 层过滤器是颜色探测器和一些是边缘检测器。这是预期的颜色是有用的它们确定图像对之间的对应关系的提示。序列被用于训练和验证 2。在训练集中的图像总数为 20501。

里程计的数据被用来计算从汽车中获得的图像对间的变换。假定相机指向的方向是 z 轴，图像平面为 XY 轴。X 轴和 Y 轴是水平和垂直图像平面的方向。在 KITTI 数据库中作为显著的相机转换是由于转换沿 Z 轴或 X 轴或旋转的 Y 轴，只有这些三个维度被用来表示相机转换。在图像对之间预测转换的任务被提出作为一个分类问题。相机转换的三个维度被划分为 20 个均匀间隔的空间。被选中的训练图像对

的帧，在最多正负 7 帧的差别间，以确保在任何给定的图像对中有一个合理的重叠。基于 SFA 的训练，图像帧对被分离最多±7 帧就被认为在时间上接近彼此。

SCNN 被训练用来预测从 227*227 像素大小的图像对区域提取大小为 370*1226 像素的图像的变换。为每个图像对，随机选择裁剪图像区域的坐标。图 1 说明典型的图像剪切。

4.2 SF 数据库

SF 数据库提供了在 136K 对图像（由 17357 独特的图像组成）之间的转换。此数据集使用谷歌街景。130K 图像对被用于训练和 6K 对被用于验证。就像 KITTI，在预测图像转化的任务被作为一个分类问题。不像 KITTI，重要的图像变换以转换的六个维度被发现（即 3 个欧拉角和 3 个转换）。因为，这是不合理的期望
这种视觉特征可以用来推断大的图像转换，在 [-30◦, 30◦]的旋转被分级为 10 个均匀间隔的空间和 2 个被用来旋转大于或小于 30◦和-30◦的额外的空间。三个转换分别放进 10 个均匀间隔空间中。图像缩放大小为 360 × 480，区域大小为 227 × 227 的图像被用来训练 SCNN。

4.3 网络体系结构

BCNN 紧随前五的结构 AlexNe 层：c96-p-c256-p-c384-c384-c256-p. TCNN 的结构是：c256-c128-f500-d-op. 在 TCNN 中的卷积过滤器是 3×3 大小的空间。这个网络被 60K 迭代与 128 批量训练。初始学习率设置为 0.001，并为每一个 20K 迭代降低的两个因素。

我们长期使用的在 KITTI 和 SF 数据库中的运动网络训练分别作为 KITTI 网络和 SF 网络在 KITTI 和 SFA 上的网络训练被叫做 KITTI-SFA 网络。图 3 显示了 KITTI 和 SF 网络的第 1 层过滤器。大部分的第 1 层过滤器是彩色探测器，而有些则是边缘检测器。由于颜色是一个确定对应相近的一个视频序列的帧的有用的线索，知道颜色探测器作为第 1 层过滤器是不足为奇的。对于 SF 网络而言，边缘检测过滤器的分数是更高的。这也不奇怪，因为在 SF 数据库中的高比例的图象中包含了结构化的对象，如建筑物和汽车。

5. 评估基于运动的学习

为了评估所提出的方法的优缺点，基于运动监督的特征学习与使用类标签和基于监督的 SPF 在场景识别，内部关键点匹配，视觉里程和目标识别方面进行了比较。特征学习的最终目标是在新任务中要从仅有的几个有监督的例子中发现特征。因此对于提供目标任务只有几个标记的例子时，评价特征的质量是有意义的。结果，在每类可用于微调的例子只有 1-20 个标记时，场景和物体识别实验进行了。

在 KITTI 网络和 SF 网络（实例模型的训练采用基于监督的运动训练）

使用仅有的 20K 独特图像进行训练。与基于类标签的训练作出公正的比较，一个 AlexNet 模型是用 20K 取自 ilsvrc12 挑战训练集的图像进行训练（即每类 20 例）。这种模式被称为 AlexNet-20K。此外，在这项工作中提出的一些实验也让 AlexNet 模型训练与被命名为 AlexNet-100K 的 100k 图象和 AlexNet-1M 的 1M 图象进行了比较。

Method	1	5	10	20
AlexNet-Scratch	1.1	3.1	5.9	14.1
KITTI-SFA-Net (Slowness)	1.5	3.9	6.1	14.9
KITTI-Net (Egomotion)	**2.3**	**5.1**	**8.6**	**15.8**

表 3：在 ILSVRC-12 验证库中对物体识别任务前 5 的精度。 AlexNet-Scratch 是指一个带有 alexnet 结构的随机权重网络的初始化。对 KITTI 网络和 KITTI-SFA 网络的权重通过在 KITTI 的数据集中使用基于运动和基于监督的 SFA 获得。所有的网络以每类 1，5，10，20 个例子进行调整。KITTI 网明显优于 AlexNet-Scratch 和 KITTI-SFA 网络。

5.1 场景识别的评估

SUN 数据库由 397 个室内/室外场景类别组成，用于评估场景识别性能。此数据库提供了每类 5 和 20 幅训练图像和每类 50 幅图象的一个标准测试库的 10 标准分割由于运行 10 个实验的时间限制，我们评估的性能，使用只有 3 个训练/测试拆分。

为了评估 CNN 的实效性，提供了不同的层次，不同的线性分类由独立的 CNN 层（即 BCNN 层该 KITTI 网络，KITTI-SFA 网络和 SF 网络）产生的特征被用来进行培训。表 2 报告了对于研究中的各个网络的识别准确率（平均超过 3 个训练/测试拆分）。 KITTI 网络优于 SF 网络，相媲美对于 AlexNet-20K。这表明，给定一个基于监督的特征学习的训练运动图像的固定的预算与使用类监督的特征训练在场景识别上几乎一样好。由 KITTI-SFA 网络和 KITTI 网络的 1-3 层（在表 2 中，简称 L1、L2、L3）计算的性能特征是可比的，而第 4 层，对 于 KITTI 网络的 5 个特点明显优于第 4 层，对于 KITTI-SFANet 的 5 个特点。这表明，基于运动的训练在于更高层次的学习特征，而 SFA 的训练只在于学习低级别的特征。

KITTI 网络的性能优于 GIST，GIST 是专门用于开发场景分类的，但优于带有空间锥匹配内核的 Dense SIFT。KITTI 网络利用有限的可视化数据（20kframes）进行训练，这些数据含有有限种类的视觉意象。KITTI 数据主要有道路，建筑物，汽车，一些行人，树木和一些植被的图像。这是一个令人惊讶的事实，网络训练数据这种小的多样性的数据在分类受过训练有着更多元化图像集的 AlexNet-20K 的室内和室外场景，是有竞争力的。我们相信伴随着更多基于学习的运动训练数据，特征学习的性能会比目前报道的数字更好。

除了第 1 层（L1）的性能外，KITTI 网络优于 SF 网络。相比于 SF 数据集（见 4.1，4.2），它是尽可能地从 KITTI 数据库中提取一个更大的图像对区域，KITTI 网络优于 SF 网络的结果是不足为奇的。因为在这个实验中，KITTI 网络被认为是优于 SF 网络的，KITTI 网络被用于本文中所描述的其他所有的实验。

5.2 在目标识别上的评估

如果为了目标识别基础训练的运动获得有用的特征，然后初始化 KITTI 网络的权值在物体识别任务上应优于随机初始化权值的网络。为了测试这个，我们训练了 CNNs 采用从 ILSVRC-2012 挑战获得的每类 1，5，10 和 20 幅图像。由于该数据集包含 1000 类，为了训练，在这些网络中可得到的培训实例的总数目分别是是 1K，5K，10K，20K 幅。对 KITTI 网络，KITTI-SFA 网络和 AlexNet-Scratch（即 CNN 随机权值初始化）的所有层为了图像分类被调整。

实验结果见表 3 结果表明基于监督（KITTI 网络）的运动明显优于基于监督（KITTI-SFA 网络）的 SFA 和 AlexNet-Scratch。正如预期的那样，基于运动的训练所提供的改进比目标任务提供较少时是更大的。这些结果表明基于训练的运动，为了物体识别，获取有用的特征。

Method	Pretrain Supervision	#Pretrain	#Finetune	L1	L2	L3	L4	L5	L6	#Finetune	L1	L2	L3	L4	L5	L6
AlexNet-1M	Class-Label	1M	5	5.3	10.5	12.1	12.5	18.0	23.6	20	11.8	22.2	25.0	26.8	33.3	37.6
AlexNet-20K		20K	5	4.9	6.3	6.6	6.3	6.6	6.7	20	8.7	12.6	12.4	11.9	12.5	12.4
KITTI-SFA-Net	Slowness	20.5K	5	4.5	5.7	6.2	3.4	0.5	-	20	8.2	11.2	12.0	7.3	1.1	-
SF-Net	Egomotion	18K	5	4.4	5.2	4.9	5.1	4.7	-	20	8.6	11.6	10.9	10.4	9.1	-
KITTI-Net		20.5K	5	4.3	6.0	5.9	5.8	6.4	-	20	7.9	12.2	12.1	11.7	12.4	-
GIST [37]	Human	-	5	6.2						20	11.6					
SPM [37]	Human	-	5	8.4						20	16.0					

表 2：为了场景识别，在 SUN 数据库中，采用基于运动和基于类标签的监督，比较神经网络预训练的准确性。这些网络 1-6 层（标记为 L1-L6）的性能被使用来自 SUN 数据库的每类 5 / 20 幅图像调整后的网络评估。KITTI 网络（即基于运动训练）的性能优于在 ImageNet 训练网络（即基于类的训练）相同数量的训练图像（即 20K）。

5.3 组内关键点匹配的评估

识别同一对象类的不同的物体的同一关键点是一个重要的视觉任务。基于监督的运动，SFA 和类标签的视觉特征为了使用关键 Pascal 数据库的注释的这项任务被评估了。

关键点匹配是通过以下方式计算：首先，地面真实物体的包围盒（GT-BBOX）从 PASCAL-VOC2012 数据库中进行提取和大小的调整（同时保持长宽比），以确保框的较小侧的长度为 227 像素。接着，为每个 GT-BBOX，计算各个 CNNs 的 2-5 层特征图。关键点匹配得分计算所有属于同一个对象类之间的 GT-BBOX 的所有对。对于给定 GT-BBOX 对，在第一个图像中，与关键点相关的特征将被用于预测在第二幅图像中相同的关键位置。在关键

点匹配中，标准化的像素距离在实际关键点和被错误预测的关键点的距离之间。关于这个程序的更多细节，在补充材料中提供。

这是很自然地期望关键点匹配精度，这将取决于在物体的观测点之间（即观点距离）的图像转换。为了对在此项任务中不同的训练方法之间的特征学习的效用进行整体的评价学习对这一任务，匹配误差被计算为一个函数的观点距离。图 4 报告了所有关键点，所有 gt-bbox 对和使用 conv-3 和 conv-4 层的特征提取的所有类的匹配误差平均值。

KITTI 网络的只用了 20K 独特帧的训练比 AlexNet-20K 和 AlexNet-100K 更有优势，比 AlexNet-1M 稍差。AlexNet 体系结构的网络随机加权初始化，令人惊讶的是其比 AlexNet-20K 表现得更好。对于这个现象一个可能的解释是通过 alexnet-20k 只有 20K 例子的特征学习只捕捉整体物体的外观，因此在关键匹配表现得很差。SIFT 是寻找图像间一致性的手工方法，对于这项任务而言，与最好的 AlexNet-1M 特征表现得一样好。（即 conv-4 特点）。KITTI 网络也明显优于 KITTISFA-Net 网络。这些结果表明，基础训练的运动特征学习在关键点匹配的任务中优于基于训练的 SFA 和内标签。

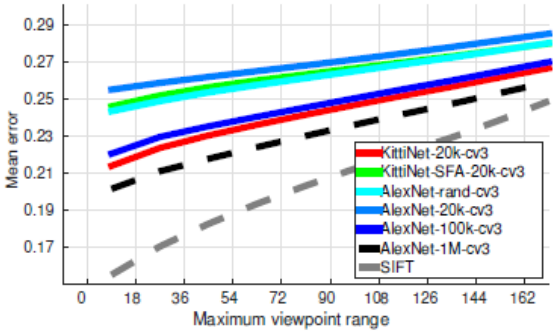
Method	Translation Acc.			Rotation Acc.		
	δX	δY	δZ	$\delta\theta_1$	$\delta\theta_2$	$\delta\theta_3$
SF-Net	40.2	**58.2**	38.4	45.0	**44.8**	40.5
KITTI-Net	**43.4**	57.9	**40.2**	**48.4**	44.0	**41.0**
AlexNet-1M	41.8	58.0	39.0	46.0	44.5	40.5

表 4:在视觉测距的任务上比较各种训练方法的准确性。

5.4 视觉里程计的评价

视觉里程计是估算在图像对之间的转换任务。在视觉里程计的任务中，所有的 KITTI 网络和 AlexNet-1M 使用了 SF 数据训练库进行了 25K 次迭代（见任务描述 4.2 节）。各种 CNNs 的性能使用了 SF 数据验证库进行了评估，结果在表 4 中。

对于这个任务而言，KITTI 网络的性能优于或相当于 alexnet-1m。在 SF 数据库中，这样的评价并不奇怪 SF 网络的一些指标优于 KITTI 网络。这个实验的结果表明，在视觉里程计的任务中，基于运动的特征学习优于基于类标签的特征学习。



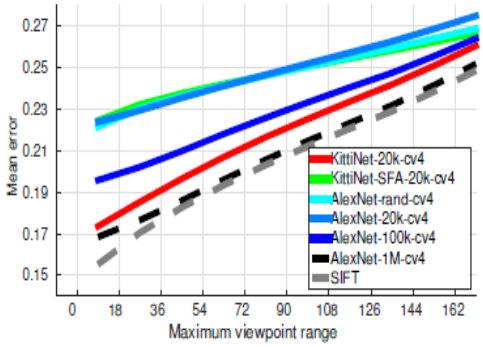


图 4:内部关键点匹配误差的函数视角距离平均超过 20PASCAL 对象，使用 conv3 层(左)和 conv4 层(右)的各种 CNNs 用于这项工作。更多细节请参阅文本。

6. 讨论

在这项工作中，我们已经表明，运动在移动设配中是一个很有用的视觉特征学习的内在监督来源。与之对应的是类标签，运动知识是容易获得的。基于 MNIST，egomotion 的特征学习优于许多以前的无监督的特征学习方法。鉴于同样的训练图像，对于场景识别任务，基于运动的特征学习学表现得几乎与以类标签为基础的特征学习一样好。此外，运动特征学习在两个数量级以上的数据（alexnet-1m）中在视觉里程计的任务中和一个数量级以上数据的类内关键点匹配任务中优于由 CNN 使用类标签监督的特征学习。另外为了证明基于运动监督的效用，这些结果还表明，基于监督的类标签的特征对于所有的视觉任务而言不是最佳的。这意味着未来的工作应该看什么样的训练对于什么任务是有用的。

我们工作的一个潜在的不足是我们对高容量深模型的培训和评估是在相对小的数据上（例如在 KITTI 数据库只有 20K 独特的图像可用）上进行的。从理论上说，通过精简的网络，我们可以获得更好地特征。例如，在我们的 MNIST 实验中，我们发现训练一个 2 层网络代替 3 层网络带来了更好的性能（表 1）。在这项工作中，我们已经有意识地选择使用标准的深度模型，因为这项工作主要的目标不是探索新的特征提取架构而是研究运动关于学习视觉表征的架构的价值，其在实际应用中表现良好。专注于探索更适合运动学习的架构的未来研究只会为这一项工作创造出强有力的示例来。对于移动机器而言，运动是容易得到的，目前还没有公开可用像 ImageNet 一样大的数据库。因此，我们无法在全方位的训练库中评估基于运动的监督的效用。在附录 B 中我们提供了一个随着训练数据的变化，基于监督的运动性能研究的初步补充材料。

在这项工作中，我们选择了第一训练模型作为一个基本的任务（即运动）然后对于目标任务调整这些模型。一个同样有趣的设置是在网上学习是移动机器有连续地获得内在的监督（如运动）和偶尔地获得外部指导信号（如类标签）。我们相信这样的训练程序很可能产生更好的特征学习。我们直觉是看到一个不同观点的相同的可能

不足够获得不同的汽车类型的实例对象，应该组合在一起。这个关于对象标签的偶尔的外部信号可以证明对于移动机器有这样的观念是有用的。同时，目前的工作被动地收集运动数据，如果有可能学习的话，去探索是否有可能获得更好的视觉表征，移动机器是否可以主动决定如何探索其环境（即主动学习[2]），那将是一件有趣的事。

致谢

这项工作有 ONR MURIN00014-14-1-0671 的部分支持。通过富布赖特科技奖学金，Pulkit Agrawal 得到部分支持。Jo˜ao Carreira 得到葡萄牙科学基金会,FCT, 格兰特 SFRH/BPD/84194/2012 的支持。我们感谢得到 NVIDIA 公司对于这项研究的 TeslaGPUs 捐赠。

补充材料

在第一作者的网站中，可以找到补充材料。

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV 2014*, pages 329–344. Springer, 2014.

[2] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

[3] H. Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[4] Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1, 2012.

[5] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, pages 168–181. Springer, 2010.

[6] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4-5):291–294, 1988.

[7] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, pages 737–744, 2011.

[8] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.

[9] J. E. Cutting. *Perception with an eye for motion*, volume 177.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[11] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

[13] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[14] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, and Y. LeCun. Unsupervised feature learning from temporal data. *arXiv preprint arXiv:1504.02518*, 2015.

[15] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, J. Han, B. Flepp, U. Muller, and Y. LeCun. Online learning for offroad robots: Using spatial label propagation to learn long-range traversability. In *RSS*, volume 11, page 32, 2007.

[16] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming

auto-encoders. In *ICANN*, pages 44 - 51. Springer, 2011.
[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675 - 678. ACM, 2014.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097 - 1105, 2012.
[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169 - 2178. IEEE, 2006.
[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541 - 551, 1989.
[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278 - 2324, 1998.
[22] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, pages 609 - 616. ACM, 2009.
[23] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150 - 1157. IEEE, 1999.
[24] R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 22(6):1473 - 1492, 2010.
[25] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, pages 737 - 744. ACM, 2009.
[26] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23 - 36, 2006.
[27] B. A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607 - 609, 1996.
[28] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. Technical report, DTIC Document, 1989.
[29] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, pages 1 - 8. IEEE, 2007.

[30] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.
[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.
[32] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *AISTATS*, pages 448－455, 2009.
[33] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568－576, 2014.
[34] S. Soatto. Visual scene representations: Sufficiency, minimality, invariance and approximations. *arXiv preprint arXiv:1411.7676*, 2014.
[35] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 41－48. IEEE, 2014.
[36] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715－770, 2002.
[37] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485－3492. IEEE, 2010.