一种用于实时六自由度摄像机重新定位的卷积网络

arXiv:1505.07427v4 [cs。简历]2016 年 2 月 18 日

亚历克斯·肯德尔·马修·格里姆斯

剑桥大学

Roberto Cipolla

agk34，mkg30，rc10001 @cam.ac.uk

Figure 1: **PoseNet: Convolutional neural network monocular camera relocalization.** Relocalization results for an input image (top), the predicted camera pose of a visual reconstruction (middle), shown again overlaid in red on the original image (bottom). Our system relocalizes to within approximately $2m$ and $6°$ for large outdoor scenes spanning $50,000m^2$. For an online demonstration, please see our project webpage: `mi.eng.cam.ac.uk/projects/relocalisation/`

摘要

1.介绍

推断你在哪里，或者定位，对于移动机器人、导航和增强现实来说至关重要。本文通过引入一种新的重新定位算法来解决丢失或被绑架的机器人问题。我们提出的系统，PoseNet，拍摄一张 224×224 的 RGB 图像，并相对于场景回归相机的 6 自由度姿态。图 1 展示了一些例子。该算法很简单，因为它由一个端到端训练的卷积神经网络(convnet)组成，用于回归摄像机的方向和位置。它实时运行，运行时间为 5 毫秒，对于大规模室外场景(覆盖地面面积高达 50，000 米)，可获得大约 2 米和 6 度的精确度。

我们的主要贡献是深度卷积神经网络相机姿态回归器。我们引入两种新颖的技术来实现这一点。我们利用非常大规模的分类数据集从识别到重新定位的转换学习。此外，我们使用运动结构从场景的视频中自动生成训练标签(相机姿态)。这减少了创建标记视频数据集的人工劳动，只需记录

我们提出了一个鲁棒的实时单目六自由度再定位系统。我们的系统训练一个卷积神经网络，以端到端的方式从单个 RGB 图像中回归出六自由度凸轮时代的姿态，而不需要额外的工程或图形优化。该算法可以在室内和室外实时运行，每帧耗时 5 毫秒。对于大规模室外场景，它获得大约 2m 和 6♀精度，而对于室

内场景，它获得 0.5m 和 10♀精度。这是使用有效的 23 层深层 convnet 实现的，表明 conv net 可用于解决复杂的图像平面外回归问题。通过利用从大规模分类数据中的转移学习，这成为可能。我们证明了点匹配网络定位于高层次的特征，并且对于困难的光照、运动模糊和不同的基于点的 SIFT 调节失败的相机内部结构是鲁棒的。此外，我们还展示了所产生的姿势特征如何推广到其他场景，从而使我们只需几十个训练示例就可以回归姿势。

视频。

我们的第二个主要贡献是理解这个 convnet 生成的表示。我们表明，该系统学习计算特征向量，这些向量很容易映射到姿态，并且通过几个额外的训练样本也可以推广到看不见的场景。

基于外观的重新定位已经取得了成功，[4，23]在一组有限的、离散的位置标签中粗略地定位了摄像机，将姿态估计留给了一个单独的系统。本文提出了一种直接从表象计算连续位姿的方法。场景可能包括多个对象，不需要在一致的条件下查看。例如，场景可以包括动态对象，如人和汽车，或者经历变化的天气条件。

同步定位与地图创建(SLAM)是解决这一问题的传统方法。我们引入了一个新的定位框架，该框架消除了典型 SLAM 管道面临的几个问题，例如需要存储密集间隔的关键帧，需要维护基于外观的定位和基于地标的姿态估计的分离机制，以及需要建立帧到帧的特征对应。我们通过将单目图像映射到一个对有害变量具有鲁棒性的高维表示来做到这一点。我们的经验表明，这种表示是一种姿态的平滑变化的(一对一)函数，允许我们直接从图像中回归姿态，而不需要跟踪。

训练卷积网络通常依赖于非常大的标记图像数据集，组装成本很高。例子包括 ImageNet [5]和 Places [29]数据集，分别有 1400 万和 700 万手动标记的图像。我们采用两种技术来克服这一限制:

一种自动标记数据的方法，使用运动结构生成相机姿态的大型回归数据集

转移学习，在大量图像识别数据集上训练预先训练为分类器的姿势回归器。与从头开始训练相比，即使训练集非常稀疏，这种方法也能在更短的时间内收敛到更低的误差。

2.相关著作

通常有两种本地化方法:基于 met-ric 和基于外观。米制 SLAM 通过集中创建稀疏的[或密集的来定位移动机器人

[环境地图。给定一个好的初始姿态估计，公制 SLAM 估计相机的连续姿态。基于外观的定位通过在有限数量的离散位置中分类场景来提供这种粗略估计。可扩展的基于外观的本地化-

已经提出了诸如[4]之类的方法，该方法在一个词袋方法中使用 SIFT 特征[15]来概率地识别以前观看过的风景。Convnets 还被用来将一个场景划分为几个位置标签中的一个([23)。我们的方法结合了这些方法的优点:它不需要一个初始的姿态估计，并产生一个连续的姿态。注意，我们不构建地图，而是训练神经网络，其大小不同于地图，不需要与场景大小成线性比例的内存(见图 13)。

我们的工作最接近于[20 中提出的场景坐标回归森林的再本地化。该算法使用深度图像来创建场景坐标标签，该标签将每个像素从相机坐标映射到全局场景坐标。然后用它训练一个回归森林来回归这些标签并定位摄像机。然而，与我们的方法不同，该算法仅限于生成场景坐标标签的 RGB-D 图像，实际上限制了其在室内场景中的使用。

以前的研究，如[27，14，9，3]也使用了类似 SIFT 的基于点的特征来匹配和定位地标。然而，这些方法需要大量的特征数据库和有效的检索方法。一种使用这些点特征的方法是从运动构造(SfM)[28，1，22]，我们在这里使用它作为离线工具来自动标记具有相机姿态的视频帧。我们使用[8]来生成我们重新定位结果的密集可视化。

尽管卷积神经网络能够对时空数据进行分类，但它们只是刚刚开始被用于回归。他们推进了物体探测的技术水平[24]和人体姿势回归[25]。然而，这些限制了他们的回归目标位于二维图像平面。在这里，我们演示了回归包括深度和平面外旋转在内的全六自由度相机姿态变换。此外，我们表明我们能够学习回归，而不是成为一个非常精细的分辨率分类器。

已经表明，在分类问题上训练的 convnet 表示法可以很好地推广到其他任务([18，17，2，6)。我们表明你可以将这些分类表示应用到 6 自由度回归问题中。使用这些预先学习的表示允许 convnets 在较小的数据集上使用而不会过度拟合。

3.相机姿态深度回归模型

在这一节中，我们描述了卷积神经网络(convnet)，我们训练该网络直接从单目图像中估计相机姿态，即，我们的网络输出一个姿态向量 p，由 3D 相机位置 x 和四元数 q 表示的方向给出:

p = [x，q]

(1)

相对于任意全局参考系定义姿态 p。我们选择四元数作为我们的方向表示，因为任意的 4-D 值很容易通过将它们归一化为单位长度来映射到理想的旋转。这是一个比旋转矩阵所需的正交化更简单的过程。

3.1 .同时学习位置和方向

为了回归姿态，我们使用随机梯度下降和下列目标损失函数训练欧几里德损失转换网络:

损耗(I)= kx x k2+β

ˇq q k q k 2

(2)

其中 β 是一个比例因子，用于保持位置和方向误差的期望值大致相等。

这组旋转位于四元数空间中的单位球面上。然而，欧几里德损失函数并不努力将 q 保持在单位球面上。然而，我们发现，在训练期间，q 变得足够接近于 q，使得球面距离和欧几里德距离之间的差别变得不明显。为了简单起见，为了避免用不必要的约束来妨碍优化，我们选择忽略球形约束。

我们发现，训练单个网络分别回归位置和方向的表现，与用全 6 自由度姿势拉贝尔训练时相比，表现不佳(图 2)。如果只有位置或方向信息，convnet 就不能有效地确定表示相机姿态的函数。我们还尝试将网络向下分成两个独立的部分，以回归位置和方向。然而，我们发现它也不太有效，原因相似:分成不同的位置和方向回归，否认了每个人从位置中排除方向的必要信息，反之亦然。

在我们的损失函数(2)中，方向和平移损失之间必须达到平衡 β(图 2)。它们是高度耦合的，因为它们是从相同的模型权重回归的。我们观察到最佳 β 由训练结束时的位置和方向的预期误差之比给出，而不是由开始时给出。我们发现室外场景的 β 值更大，因为位置误差相对更大。根据这种直觉，我们使用网格搜索微调 β。室内场景在 120 到 750 之间，室外场景在 250 到 2000 之间。

我们发现随机初始化最终位置回归层非常重要，这样每个位置维度对应的权重范数与该维度的空间范围成正比。

分类问题有一个每个类别的训练例子。这对于回归是不可能的，因为

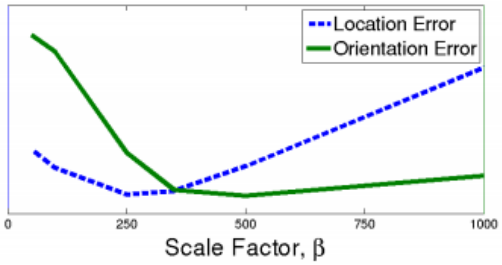


图 2:在一个单一的 convnet 上位置和方向的相对性能，带有一系列室内场景的比例因子，象棋。这表明使用最佳比例因子进行学习可以使 convnet 发现更精确的姿态函数。

输出是连续和无限的。此外，已经用于回归的其他数据集在非常大的数据集上运行[25，19]。为了使局部化回归处理有限的数据，我们通过对这些数据集的权重进行预处理，利用从这些大型分类数据集中获得的强大表示。

3.2 .体系结构

在本文的实验中，我们使用了一种先进的深层神经网络结构进行分类，谷歌[24]，作为开发我们的姿态反馈网络的基础。GoogLeNet 是一个 22 层的卷积网络，有六个“初始模块”和两个额外的中间分类器，在测试时被丢弃。我们的模型是有 23 层的 GoogLeNet 的一个稍加修改的版本(只计算有可训练参数的层)。我们对谷歌网站做了如下修改:

用仿射遗憾替换所有三个软最大分类器。移除 softmax 层，并修改每个最终完全连接的层，以输出表示位置(3)和方向(4)的 7 维姿态向量。

在特征尺寸为 2048 的最终回归之前，插入另一个完全连接的层。这是为了形成一个局部化的特征向量，然后可以对其进行探索以进行推广。

在测试时，我们还将四元数方向向量归一化为单位长度。

我们重新调整了输入图像的比例，使其最小尺寸为 256 像素，然后裁剪为 224x224 像素，放入 GoogLeNet convnet。convnet 是在随机作物(不影响相机姿态)上训练的。在测试时，我们用输入图像的单个中心裁剪和 128 个均匀间隔的裁剪对其进行评估，平均得到的姿态向量。通过并行的图形处理器处理，每幅图像的计算时间从 5 毫秒增加到 95 毫秒。

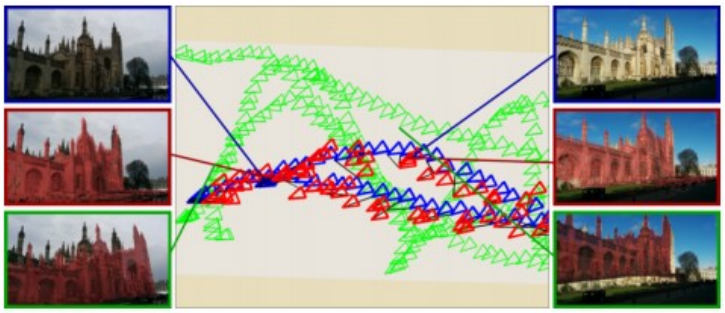


图 3:国王学院一系列训练(绿色)和测试(蓝色)摄像机的放大图。我们用红色显示每个测试帧的预测摄像机姿态。这些图像显示了测试图像(上图)，我们的 convnet 预测视图用红色覆盖在输入图像上(中图)，最近邻训练图像用红色覆盖在输入图像上(下图)。这表明我们的系统可以在训练帧之间的空间中有效地插值相机姿态。

为了训练和测试，我们尝试在裁剪前将原始图像重新缩放到不同的大小。放大输入相当于在一侧向下采样到 256 像素之前裁剪输入。这增加了输入像素的空间分辨率。我们发现这并没有提高本地化性能，这表明上下文和视野比重新本地化的解决方案更重要。

PoseNet 模型是使用[10]咖啡馆图书馆实现的。它使用随机梯度去噪进行训练，基本学习率为 10-5，每 80 个时期减少 90%，动量为 0.9。使用双图形处理器卡的一半(英伟达 Titan Black)，训练用了一个小时，批量为 75。由于时间的原因，我们没有探索多图形处理器的训练，尽管有理由期待使用双倍的吞吐量和内存会有更好的结果。我们为每个场景减去一个单独的图像平均值，因为我们发现这可以提高实验性能。

4.资料组

深度学习在大型数据集上表现非常好，但是产生这些数据集通常是昂贵的或者非常劳动密集型的。我们通过利用运动的结构来自主生成训练标签(相机姿态)来克服这个问题。这减少了仅仅记录每个场景的视频的人工劳动。

在这篇文章中，我们发布了一个户外城市定位数据集，剑桥地标，有 5 个场景。这个新的数据集为在大规模户外城市环境中训练和测试姿势回归算法提供了数据。相机姿态的鸟瞰图在图 4 中示出，并进一步显示

此处提供了 PoseNet 代码和数据集:

mi.eng.cam.ac.uk/projects/relocalisation/

在表 6 中可以找到尾部。出现了大量的城市杂波，如行人和车辆，数据来自许多不同的时间点，代表不同的照明和天气条件。训练和测试图像取自不同的行走路径，而不是从同一轨迹采样，使得回归具有挑战性(见图 3)。我们发布这个数据集供公众使用，并希望随着项目的进展向这个数据集添加场景。

数据集是使用运动技术[28]的结构生成的，我们将它用作本文的地面真实测量。行人使用谷歌 LG Nexus 5 智能手机在每个场景周围拍摄高清视频。该视频在 2Hz 时进行二次采样，生成图像输入到 SfM 流水线。每个摄像机位置之间有大约 1 米的间距。

为了在室内场景上进行测试，我们使用公共可用的 7 个场景数据集[20]，场景如图 5 所示。该数据集包含相机高度的显著变化，是为 RGB-D 重新定位而设计的。对于使用 SIFT 特征的纯视觉再定位来说，这是非常具有挑战性的，因为它包含许多模糊的无纹理特征。

5.实验

我们在表 6 中展示了 PoseNet 能够有效地定位室内 7 个场景数据集和室外 Cam-bridge Landmarks 数据集。为了验证 convnet 的回归姿态超出了训练样本，我们展示了从定位 convnet 产生的特征向量中找到训练数据中最近邻表示的性能。当我们的性能超过这个值时，我们得出结论，convnet 能够成功地在训练示例之外回归姿态(见图 3)。我们还将我们的算法与[20]的 RGB-D 分数森林算法进行了比较。

图 7 示出了两个室内场景和两个室外场景的定位误差的累积直方图。我们注意到，尽管 SCoRe 森林通常更精确，但它需要深度信息，并且使用更高分辨率的管理器。室内数据集包含许多模糊和无纹理的特征，这使得没有这种深度模态的重新定位非常困难。我们注意到我们的方法经常定位最困难的测试帧，在 95%以上，在所有场景中比 SCoRe 更精确。我们还观察到，密集裁剪只能适度提高性能。它在行人和汽车等非常杂乱的场景中最为重要，例如国王学院、商店门面和圣玛丽教堂。

我们用来自黄昏、雨、雾、夜和运动模糊的额外图像以及具有未知内在结构的不同相机，在数据集中测试了该方法的鲁棒性。图 8 示出了转换网络的总体结构

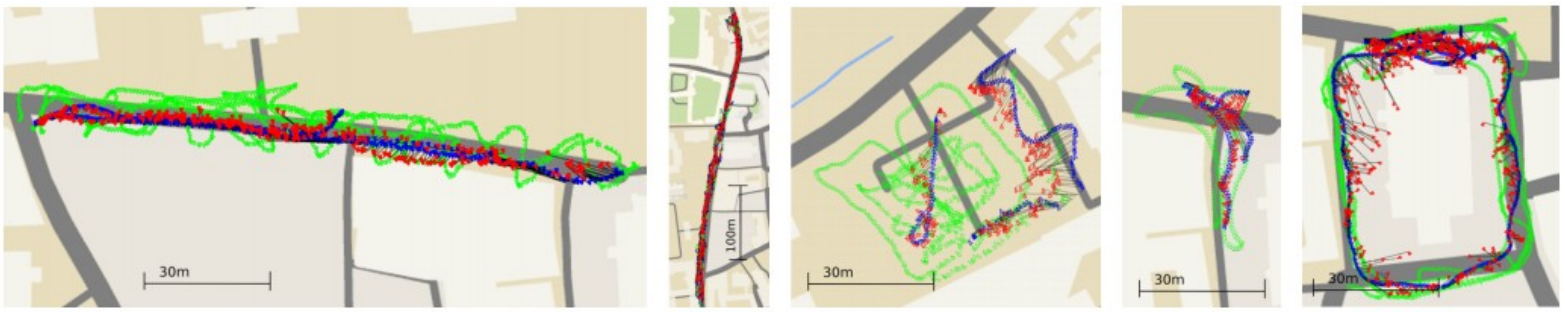


Figure 4: **Map of dataset** showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.

图 5: 7 从左到右的场景数据集示例图像；象棋，火，头，办公室，南瓜，红色厨房和楼梯。

#帧空间分数森林距离。去 Conv。

场景训练测试范围(m)(使用 RGB-D)最近邻点密集点

国王学院 1220 343 140 x 40m 米北/南 3.34 米，5.92♀1.92 米，5.40♀1.66 米，4.86♀

街道 3015 2923 500 x 100m 北/北 1.95m，9.02♀3.67m，6.50♀2.96m，6.00♀

旧医院 895 182 50 x 40m 米北/东 5.38 米，9.02♀2.31 米，5.38♀2.62 米，4.90♀

车间正面 231 103 35 x 25m 米 N/A 2.10 米，10.4♀1.46 米，8.08♀1.41 米，7.18♀

圣玛丽教堂 1487 530 80 x 60m 米北/东 4.48 米，11.3♀2.65 米，8.48♀2.45 米，7.96♀

国际象棋 4000 2000 3 x 2 x 1m 米 0.03 米，0.66♀0.41 米，11.2♀0.32 米，8.12♀0.32 米，6.60♀

消防 2000 2000 2.5 x 1 x 1m 米 0.05 米，1.50♀0.54 米，15.5♀0.47 米，14.4♀0.47 米，14.0♀

水头 1000 1000 2 x 0.5 x 1m 0.06m，5.50♀0.28m，14.0♀0.29m，12.0♀0.30m，12.2♀

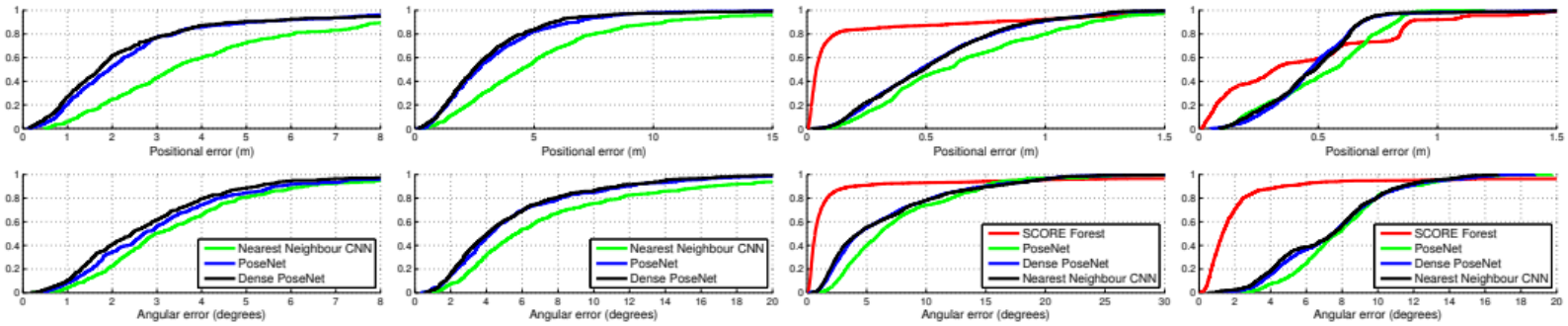
办公室 6000 4000 2.5 x 2 x 1.5m 0.04m，0.78♀0.49m，12.0♀0.48m，7.68♀0.48m，7.24♀

南瓜 4000 2000 2.5 x 2 x 1m 米 0.04 米，0.68♀0.58 米，12.1♀0.47 米，8.42♀0.49 米，8.12♀

红色厨房 7000 5000 4 x 3 x 1.5m 0.04m，0.76♀0.58m，11.3♀0.59m，8.64♀0.58m，8.34♀

楼梯 2000 1000 2.5 x 2 x 1.5m 0.32m，1.32♀0.56m，15.4♀0.47m，13.8♀0.48m，13.1♀

图 6:数据集细节和结果。我们在所有场景中展示了 PoseNet 的中值性能，对单个 224x224 中心作物和 128 个均匀分布的密集作物进行了评估。为了比较，我们绘制了使用深度的[森林 20]的结果，因此在室外场景中失败。该系统以大得多的分辨率回归输入图像的像素世界坐标。这需要密集的深度图进行训练，并需要额外的 RANSAC 步骤来确定相机的姿态。此外，我们还比较匹配最近邻特征向量表示。这表明我们的回归 PoseNet 比分类器性能更好。



国王学院

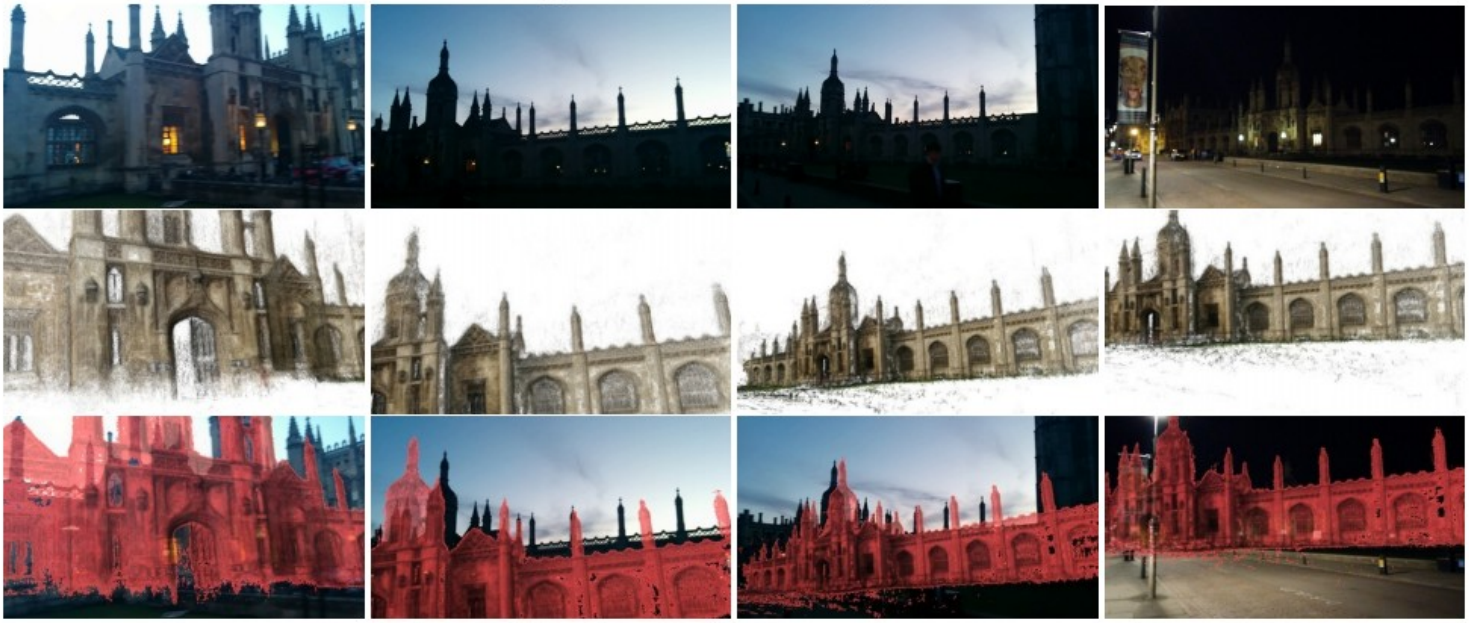
南瓜

楼梯

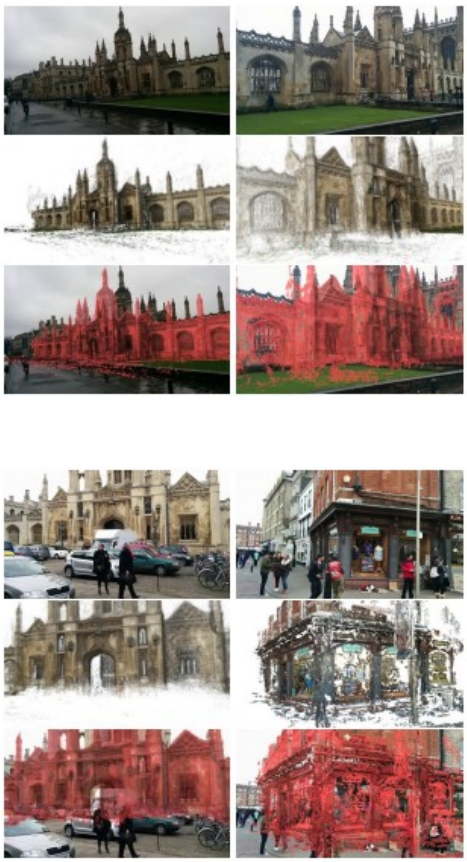
圣玛丽教堂

图 7:本地化性能。这些数字显示了我们的定位精度的位置和方向，作为整个测试集的累积误差。回归 convnet 的性能优于最近邻特征匹配，这表明我们回归比训练给出的更好的分辨率结果。与 RGB-D SCoRe Forest 方法相比，我们的方法是有竞争力的，但是比更昂贵的深度方法性能更好。我们的方法在最难的几个帧上表现得更好，超过了第 95 个百分点，我们的最差误差低于 SCoRe 方法的最差误差。

(a) Relocalization with increasing levels of motion blur. The system is able to recognize the pose as high level features such as the contour outline still exist. Blurring the landmark increases apparent contour size and the system believes it is closer.



(b) Relocalization under difficult dusk and night lighting conditions. In the dusk sequences, the landmark is silhouetted against the backdrop however again the convnet seems to recognize the contours and estimate pose.



(c)根据不同的天气条件重新定位。PoseNet 能够有效地估计雾和雨中的姿态。

(d)与重要人物、车辆和其他动态物体重新定位。

(e)用未知的相机时代的内在因素重新定位:焦距为 45 毫米(左)的单反相机，焦距为 35 毫米(右)的 iPhone 4S，而数据集的相机焦距为 30 毫米。

图 8:对现实生活挑战的鲁棒性。在示例(a-c)中，使用基于点的技术(如 SIFT)进行配准失败，因此地面真实测量不可用。在训练过程中，没有发现这些类型的挑战。由于 convnets 能够理解物体和轮廓，因此它们仍然能够成功地从轮廓示例(b)中的建筑物轮廓或甚至在极端运动模糊(a)下估计姿态。这些准不变性中的许多都是通过对场景数据集进行预处理来增强的。

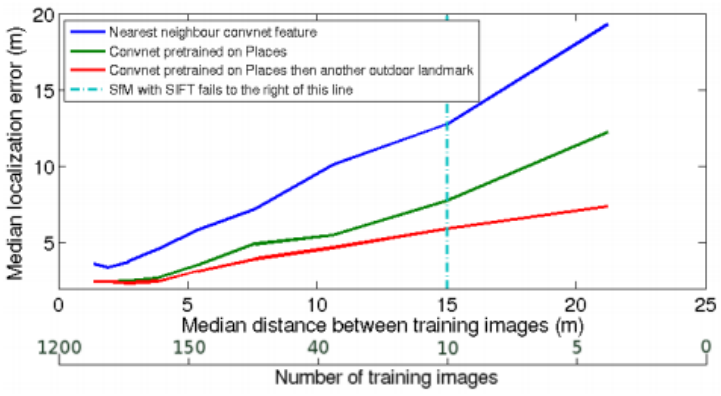


图 9:国王学院场景对降低训练基线的鲁棒性。由于使用的训练样本越来越少，我们的系统在性能上表现出适度的下降。

ally 很好地应对了这些挑战。在所有这些情况下，带有 SIFT 的 SfM 都失败了，因此我们不能生成地面真实相机姿态，但是我们通过从预测的相机姿态观察 3D 重建并将其叠加到输入图像上来推断精度。

5.1 .对训练图像间距的鲁棒性

我们在图 9 中证明，对于室外规模的场景，我们通过将训练图像的间距更接近于 4m 来获得很小的增益。该系统对训练图像之间很大的空间分离具有鲁棒性，即使只有几十个训练样本也能达到合理的性能。随着训练图像间距的增加，姿态精度会逐渐下降，而基于 SIFT 的 SfM 在达到某个阈值后会急剧下降，因为它需要较小的基线[15]。

5.2 .迁移学习的重要性

一般来说，convnets 需要大量的训练数据。我们回避了这个问题，从一个网络开始我们的姿势训练，这个网络是在大型数据集上预处理的，比如 Im-ageNet 和 Places。类似于已经为分类任务演示的，图 10 示出了如何在分类和复杂的回归任务之间有效地利用迁移学习。这种“迁移学习”已经在其他地方被证明用于训练分类器[18，17，2]，但是在这里我们证明了从分类到姿势回归的性质不同的任务的迁移学习。目前还不明显的是，一个经过训练输出姿态不变分类标签的网络是否适合作为姿态回归器的起点。然而，我们发现这在实践中不是问题。一种可能的解释是，为了使其输出不随姿态而变，分类器网络必须跟踪姿态，以便更好地将它的影响从身份提示中分离出来。这与我们自己的发现是一致的，即训练输出位置和方向的网络优于训练只输出位置和方向的网络

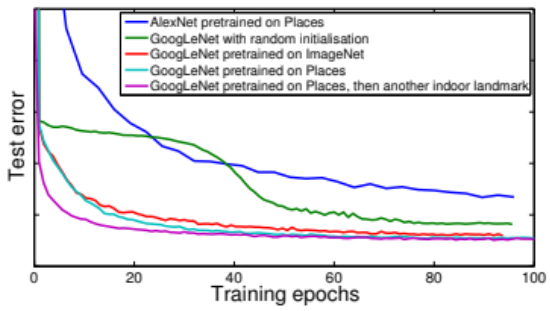


图 10:迁移学习的重要性。展示了在大型数据集上进行预训练如何提高性能和训练速度。

位置。通过保留中间表示中的方向信息，可以更好地将方向的影响排除在最终位置估计之外。转移学习不仅大大提高了训练速度，而且提高了最终成绩。

数据的相关性也很重要。在图 10 中，位置和图像网络曲线最初具有相同的性能。然而，最终 Places 预处理性能更好，因为它是与此本地化任务更相关的数据集。

5.3 .可视化与姿势相关的特征

图 11 示出了由 PoseNet 产生的示例显著图。在[21 中使用的显著图]是损耗函数的梯度相对于像素强度的大小。这使用姿势相对于像素的敏感度作为 convnet 考虑图像不同部分的重要程度的指标。

这些结果表明，最强的响应来自更高层次的特征，如窗口和尖顶。然而，一个更令人惊讶的结果是，波塞内对大的无纹理斑块，如道路、草地和天空也非常敏感。这些无纹理面片可能比最高响应点更有信息性，因为一组像素对姿态变量的影响是该组像素上显著图值的总和。这一证据表明，网络能够从这些无纹理表面定位信息，这是基于兴趣点的特征(如 SIFT 或 SURF)所不能做到的。

最后一个观察结果是，波塞内对人和其他噪声物体的反应减弱，有效地屏蔽了它们。这些对象是动态的，convnet 已经确定它们不适合本地化。

5.4 .查看内部表示

t-SNE·[26]是一种在低维空间中嵌入高维数据的算法，试图保持欧几里得距离。正如我们在这里所做的那样，它经常被用来可视化

图 11:显著图。该图显示了叠加在输入图像上的显著图。显著性图表明，convnet 不仅利用了独特的点特征(一个 la SIFT)，而且还利用了大的无纹理的面片，这些面片对于姿势来说即使不是更有信息性，也是一样的。这与忽视动态物体(如行人)的倾向相结合，使其能够在具有挑战性的环境下表现良好。(最好以电子方式查看。)

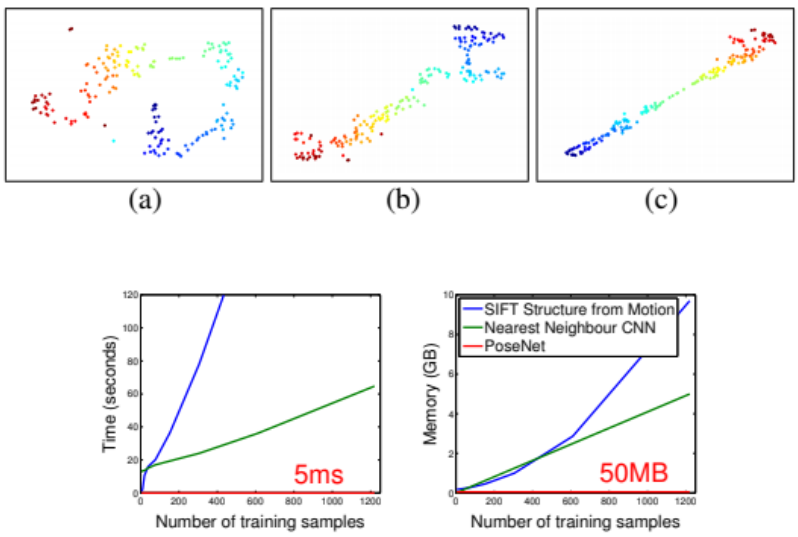


图 12:特征向量可视化。直线穿越室外场景(国王学院)的视频序列的特征向量的 t-SNE 可视化。颜色代表时间。该特征表示是从具有在地点(a)、地点然后另一个室外场景、圣玛丽教堂(b)、地点然后这个室外场景、国王学院(c)上训练的权重的 convnet 生成的。尽管(a，b)没有在这个场景中被训练，但是这些视觉化表明，通过这些表示的简单(如果非线性的话)函数来计算姿态是可能的。

图 13:实现效率。实验速度和内存使用了 convnet 回归，最近邻 convnet 特征向量和 SIFT 再定位方法。

将每个姿势摆好，与用 SIFT 进行公制定位的千兆字节和分钟进行比较。这些值与系统中训练样本的数量无关，而仪表定位用训练数据大小[28 来衡量 0(n)。为了进行比较，还显示了与 convnet 最近邻的匹配。这需要存储每个训练帧的特征向量，然后执行线性搜索以找到给定测试帧的最近邻。

二维。在图 12 中，我们将 t-SNE 应用于从行人拍摄的视频帧序列计算的特征向量。如这些图所示，特征向量是一个随姿势平滑变化的函数，并且很大程度上是一对一的。这种“姿态流形”不仅可以在其他场景上训练的网络上观察到，而且可以在没有姿态的分类图像集上训练的网络上观察到。这进一步表明，无论输出中是否表达，分类都会将预先服务的姿态信息传递到最后一层。然而，对于没有在姿态数据上训练的网络，从特征向量到姿态的映射变得更加复杂。此外，由于这个流形存在于 convnet 没有训练过的场景中，convnet 必须学习地标、几何图形和相机运动之间关系的一些一般表示。这表明从回归中产生的特征向量能够以与分类转换网相同的方式推广到其他任务。

6.结论

据我们所知，我们首次将深度卷积神经网络应用于端到端的 6 自由度摄像机姿态定位。我们已经证明，通过使用从被训练为分类器的网络的转移学习，可以避开对数百万个训练图像的需要。我们证明了这样的网络在它们的特征向量中保留了大量的姿态信息，尽管被训练来产生姿态不变的输出。我们的方法可以容忍导致基于 SIFT 的定位器急剧失效的大基线。

在未来的工作中，我们的目标是进一步使用多视图几何作为深度姿态回归器的训练数据来源，并探索该算法的概率扩展[12]。很明显，一个有限的神经网络在它可以学习定位的物理区域上有一个上限。我们把找到这个极限留给未来的工作。

5.5。系统效率

图 13 比较了 PoseNet 在现代台式计算机上的系统性能。我们的网络可扩展性很强，只需 50 兆存储重量，5 兆通信

参考

[1] S .阿加瓦尔，y .古川，n .斯纳福利，I .西蒙，b .柯勒斯，S. M .塞兹和 r .斯泽斯基。一天之内建成罗马。《奥地利通讯》，第 54(10):105-112 页，2011 年。

[2]本吉奥、库维尔和文森特。表征学习:回顾与新视角。模式分析和机器智能，美国电气和电子工程师学会学报，35(8):1798-1828，2013。

[3]佛手柑、桑娜辛哈和托雷萨尼。利用来自运动的结构来学习可分级地标分类的区别性代码本。在计算机视觉和模式识别(CVPR)，2013 年 IEEE 会议，763- 770 页。IEEE，2013。

[4]康明斯和纽曼。外观空间中的概率局部化和映射。国际机器人研究杂志，27(6):647-665，2008。

[5]邓军，董伟，索彻，李林军，李金凯，飞飞.Imagenet:一个大规模的分层图像数据库。《计算机视觉和模式识别》，2009 年。CVPR 2009。美国电气和电子工程师学会会议，第 248-255 页。IEEE，2009。

[6]唐纳休，贾，温雅斯，霍夫曼，张，曾轶可，达雷尔.脱咖啡因咖啡:一种用于一般视觉识别的深层卷积激活特性。arXiv 预印本 arXiv:1310.1531，2013。

[7]恩格尔，斯科普斯和德·克雷默斯。大规模直接单眼单眼单眼单眼。《计算机视觉——ECCV 2014》，第 834-849 页。斯普林格，2014 年。

[8]古川，柯勒斯，塞兹和塞斯基。面向互联网的多视角立体影像。在计算机视觉和模式识别(CVPR)，2010 年 IEEE 会议，第 1434-1441 页。IEEE，2010。

[9]郝，蔡，李，张，庞，吴.用于地标识别的 3d 视觉短语。在计算机视觉和模式识别(CVPR)，2012 年 IEEE 会议，第 3594-3601 页。IEEE，2012。

[10]贾，舍勒哈默，唐纳休，卡拉耶夫，龙，吉时克，，达雷尔.Caffe:用于快速特征嵌入的卷积结构。arXiv 预印本 arXiv:1408.5093，2014。

[11]梅斯，约翰松，罗伯特，维拉，伦纳德和德拉特。iSAM2:使用贝叶斯树的增量平滑和映射。《国际机器人研究杂志》，第 0278364911430419 页，2011 年。

[12] A .肯德尔和 R .西波拉。相机再定位深度学习中的建模不确定性。arXiv 预印本 arXiv:1509.05909，2015。

[13]克莱因和默里。小型 ar 工作空间的并行跟踪和映射。在混合和增强现实，2007 年。ISMAR 2007。第六届美国电气工程师学会和美国机械工程师学会国际研讨会，第 225-234 页。IEEE，2007。

[14]李亚男，斯纳利，胡特罗彻，福安。使用 3d 点云的全球姿态估计。《计算机视觉——ECCV 2012》，第 15-29 页。斯普林格，2012 年。

[·洛维。尺度不变关键点的独特图像特征。国际计算机视觉杂志，60(2):91-110，2004。

[16]纽科姆、洛夫格罗夫和戴维森。DTAM:实时密集跟踪和绘图。在计算机视觉(ICCV)，2011 年 IEEE 国际会议，第 2320-2327 页。IEEE，2011。

[17]奥库布、波图、拉普捷夫和西维克。使用进化神经网络学习和传递中级图像表示。在计算机视觉和模式识别(CVPR)，2014 年 IEEE 会议，第 1717-1724 页。IEEE，2014。

[18]拉扎维安，阿齐兹普尔，苏利文和卡尔松。Cnn 的特色是现成的:令人震惊的识别基线。在计算机视觉和模式识别工作室(CVPRW)，2014 年美国电气和电子工程师学会会议，第 512-519 页。IEEE，2014。

[19]塞曼内，艾根，张，马蒂厄，弗格斯，乐村.过度进食:使用卷积网络的综合识别、定位和检测。arXiv 预印本 arXiv:1312.6229，2013。

[20] J .肖特顿，b .格洛克，c .扎克，s .伊扎迪，A. Criminisi，和 a .菲茨基本。场景坐标回归森林用于三维图像中的凸轮时代再定位。在计算机视觉和模式识别(CVPR)，2013 年 IEEE 会议，2930-2937 页。IEEE，2013。

[21]西蒙扬，韦达迪和塞塞曼。卷积网络的深层:可视化图像分类模型和显著图。arXiv 预印本 arXiv:1312.6034，2013。

[22]斯纳弗利、塞兹和斯泽斯基。照片旅游:探索 3d 照片收藏。在美国计算机学会图形交易(TOG)，第 25 卷，第 835-846 页。奥地利中心，2006 年。

[23]北桑德海夫、弗·达约布、什拉济、布·奥克罗夫特和米尔福德。convnet 特征在位置识别中的性能研究。arXiv 预印本 arXiv:1501.04158，2015。

[24]塞格迪、刘炜、贾、塞马奈、里德、安古洛夫、洱海、范豪克和诺维奇.盘旋而下。arXiv 预印本 arXiv:1409.4842，2014。

[25]托舍夫和塞格迪。深度姿势:通过深度神经网络进行人体姿势估计。在计算机视觉和模式识别(CVPR)，2014 年 IEEE 会议，第 1653-1660 页。IEEE，2014。

[26]范德马滕和韩丁。使用 t-SNE 可视化数据。机器学习研究杂志，9(2579- 2605):85，2008。

[27]王俊杰，查海平，和 r .西波拉.通过对尺度不变特征进行索引，从粗到细进行基于视觉的定位。系统、人和控制论，第二部分:控制论，美国电气和电子工程师学会学报，36(2):413-422，2006。

28]吴。从运动走向线性时间增量结构。3D 视觉-3DV 2013，2013 年国际会议，第 127-134 页。IEEE，2013。

[29]周勃，拉普达里扎，肖，托拉尔巴，奥利瓦.使用地点数据库学习场景识别的深层特征。神经信息处理系统进展，第 487-495 页，2014。