基于单目图像深度特征压缩的视觉识别与定位

1.4.1 总体思想

如何在动态环境中快速、准确地自主定位机器人，是机器人可靠导航的首要问题。视觉特征提取是机器人位置识别的一个基本问题，其中采用的视觉定位特征是确定定位性能的关键；因此，有许多的研究学者聚焦在计算机视觉特征表示研究中。

1.4.1.1

要表达一个变化显著的场景是一个很大的挑战，如图1.4.1所示。最近的文献提出了各种方法来解决这个领域的挑战[20] [6] [22] [19] [21] [5] [18]。手动提取特征之局部特征(local features)，近年仍是来研究的一个热点。局部特征是指一些能够稳定出现并且具有良好的可区分性的特征点。这在计算机视觉和机器学习社区的几乎所有重要任务在如何手动提取特征上，以发现和描述从图像中提取的特征，如 SURF 、ORB 、BRIEF 、SIFT 、Harris 。SIFT 和 HOG中都取得了最先进的性能 。这样在物体不完全受到遮挡的情况下，一些局部特征依然稳定存在，以代表这个物体(甚至这幅图像)，方便接下来的分析。一方面，局部特征对姿态变化不太敏感，仍然恢复一些重要信息，甚至部分关键点被遮蔽或移动，可以实现图像检索的快速计算。一方面说，如果我们用这些稳定出现的点来代替整幅图像，可以大大降低图像原有携带的大量信息，起到 减少计算量的作用。另一方面，当物体受到干扰时，一些冗余的信息(比如颜色变化平缓的部分和直线)即使被遮挡了，我们依然能够从未被遮挡的特征点上还 原重要的信息。局部特征无法在如图1.4.1等复杂环境下工作。如果用户对整个图像的整体感兴趣，而不是前景本身感兴趣的话，全局特征用来描述总是比较合适的。但是无法分辨出前景和背景却是全局特征本身就有的劣势，特别是在我们关注的对象受到遮挡等影响的时候，全局特征很有可能就被破坏掉了。

(a) Rainy night

(b) Rainy daytime

(c) Shadows of trees and buildings and others

(d) Moving objects

(e) Dusk

(f) Light outsides

图 1.4.1

1.4.1.2 非手动提取特征

如何解决手动提取特征的局限以提出更有效的方法呢?众所周知，非手工制作的特征能够通过深度卷积神经网络 (DCNN) 从数百万标记图像中自动学习鉴别特征。在2012年的AlexNet大规模视觉识别挑战赛（ILSVRC）上得到了令人难以置信的准确率[10]，ConvNets已经被证明优于传统的手工制作的特征[3][23][3][16]。它是用120万张标记的图像进行预训练的。根据从AlexNet中提取的特征对图像进行分类。每个单独层的输出可以作为一个全局的图像描述符。我们还可以根据这些特征对图像进行匹配，然后对机器人进行定位。来自CNNs中间层的特征可以更有效地消除数据集的偏差。[28]比较了不同层级特征的性能。他们的结果表明，ConvNet层次结构中中间层的特征对一天中的时间、季节或天气条件引起的外观变化表现出鲁棒性。来自Conv3层的特征在外貌变化方面表现得相当好。表1列出了AlexNet ConvNets中不同层的向量维度。

[28]证明了 Conv3 层的特征在外貌变化方面的表现相当好. 此外，[28]还指出fc6和fc7在视角变化方面优于其余层。然而，当外观变化时，fc6和fc7完全失败。Conv3的维度为64896，即一幅图像显示为64896维度的矢量。在线定位将连续接收来自摄像头的图像。毫无疑问，大量的向量数学运算是很耗时的。DCNN中包含的不同特征最初是以浮动格式返回的。为了方便后续的二值化，[1]将这些特征投向一个标准化的8位整数格式。然后利用汉明距离对所有的二进制特征进行匹配，计算出一个匹配矩阵。他们的结果表明，对特征的压缩可以降低其描述符99.59%的重冗余度，而精度仅降低2%左右。此外，他们对特征的二值化允许使用汉明距离，这也代表了匹配位置的加速。

图像匹配是继特征提取之后的另一个挑战。机器人对世界的认识必须以地图的形式存储，并与当前的观察结果进行比较。[17]指出，根据视觉传感器的不同，以及进行何种类型的场所识别，地图框架也有所不同。可分为纯图像检索、拓扑地图和拓扑-计量地图。纯图像检索只存储环境中每个地方的外观信息，没有相关的位置信息，就像FAB-MAP[7]中使用的周柳树一样。FAB-MAP[7]描述了一种概率方法来解决图像和地图增强的匹配问题。他们使用了基于向量的描述符，如SURF与bag-of-words联合使用。本文学习了一个地方外观的生成模型。他们构建了一个Chow- Liu树[4]来捕捉视觉词的共现统计。Chow-Liu树由节点和边组成。变量之间的相互信息通过树的边的粗细来显示。图中的每一个节点对应一个由输入感官数据转换而来的词袋表示。FAB-MAP在具有挑战性的户外环境中成功地检测到了大部分的环路闭合。但[21]的结果显示，在跨季节的数据集中，OpenFABMAP2只能找到少数正确的匹配，原因是手工制作的特征描述符不可重复。论文[21]将图像匹配制定为数据关联图中的最小成本流问题，以有效利用序列信息。他们通过最小成本流定位车辆。他们的方法在动态场景中效果良好。[12]提出了一种马尔科夫半监督的聚类方法及其在拓扑图提取中的应用。至于增量映射、slam和导航任务，该方法可以进行相应的调整。

SeqSLAM[20]将图像识别问题框定为在局部邻域内寻找所有与当前图像最佳匹配的模板。它很容易实现。然而，[20]的算法很容易受到机器人速度的影响。这种约束限制了长时间定位的应用。[24]证明了如果只使用每张图像中信息量最大的特征，地方识别性能会有所提高。14]描述了一种使用自适应描述符的轻量级新型场景识别方法，该方法基于颜色特征和几何信息。[13]提出了一种使用轻量级自适应描述符的拓扑图全向视觉的场景识别方法。[11]用减少的特征集改进了地方识别。[15]利用非参数的Dirichlet分层模型，提出了识别和聚类问题的通用框架，命名为DP-Fusion。

我们提出一个新的视觉定位图像特征提取方法， 它结合 CNNs 网络特征表示的优势，在不同季节等环境条件下执行基于单目视觉的鲁棒定位，正如方法框架图(图1.4.2)所示，我们的工作过程如下:1)从 AlexNet 的 Conv3 中提取特征，并通过 IPCA 进行图像特征降维。2) 将在线图像的向量与已存数据集的向量通过余弦距离逐一匹配。通过核化方法对匹配矩阵进行归一 化，以减少因大部分在线图像匹配的数据集混乱造成的歧义。3)对匹配图像进 行图像处理，包括图像二值化、图像侵蚀等。4)设置参数，通过 RANSAC(随 机样本共识)在线寻找最佳匹配序列。

本文的研究过程如下。在第1.4.2节中，我们描述了我们的方法的细节。在第1.4.3节中，我们在Norland 数据集上做了一个动态环境下的特征提取与识别定位实验。在第1.4.4节中，我们对结果进行了分析与讨论。

1.4.2 单目图像深度特征压缩的视觉特征提取

关于地图框架，我们采用的是纯图像检索，但数据集是按照图像的入库时间顺序存储的。这样的话，我们不仅可以保证准确率，而且计算效率高。我们选择AlexNet的Conv3中的特征作为我们的整体图像描述符。Conv3的维度为64896，也就是说一张图像显示为64896维度的向量f，我们根据每张图像的位置建立视觉地图

$\{[f,l]\}_{ni}=1$。所以当前图像序列表示为$\{l\}_j=t-m+1$。高维向量导致耗时。我们考虑通过IPCA来减少维度。虽然图像描述符会损失一些信息，但它减少了因天空、地面和树木等数据集的混乱而造成的模糊匹配。在线图像的向量将通过余弦距离与数据集向量进行逐一比较。然后我们得到匹配矩阵S，其元素浮动在(0, 1)范围内。通过核方法对匹配矩阵进行归一化处理，以减少因与大部分在线图像匹配的数据集混乱而引起的歧义。然后通过合适的阈值化将其转换为二进制灰色图像。我们尝试调整参数，然后通过RANSAC在线寻找最佳匹配序列。匹配矩阵中当前图像的最佳匹配特征为f_{km+b}。那么当前图像在视觉图中的最佳匹配图像为l_{km+b}。

我们从Caffe提供的AlexNet的Conv3中提取特征作为我们的图像整体描述符。Conv3的维度为64896，即一幅图像由64896个维度的向量来表示。AlexNet ConvNets中不同层的向量尺寸列于表1中[10]。[28]给我们的结论是，层次结构中较高的层在语义上更有意义，但因此失去了对同一语义类型的场景中各个地方的分辨能力。决定我们使用哪一层很重要。来自Conv3层的特征在外观变化极端时仍表现较好。

表1

Layer	Dimensions	Layer	Dimensions
Conv1	96×55×55	Conv4	384×13×13
pool1	96×27×27	Conv5	256×13×13
Conv2	256×27×27	fc6	4096×1×1
pool2	256×13×13	fc7	4096×1×1
Conv3	384×13×13	fc8	1000×1×1

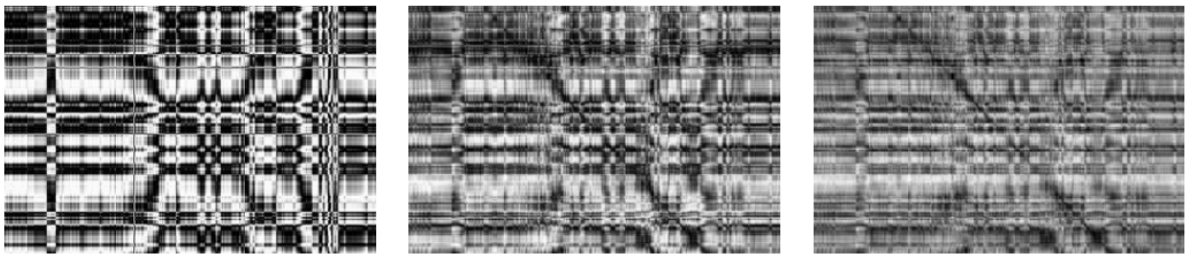
表 2

n_components	Information ratio	n_components	Information ratio
316	99%	51	93%
187	98%	44	92%
136	97%	38	91%
99	96%	33	90%
76	95%	29	89%
62	94%	25	88%

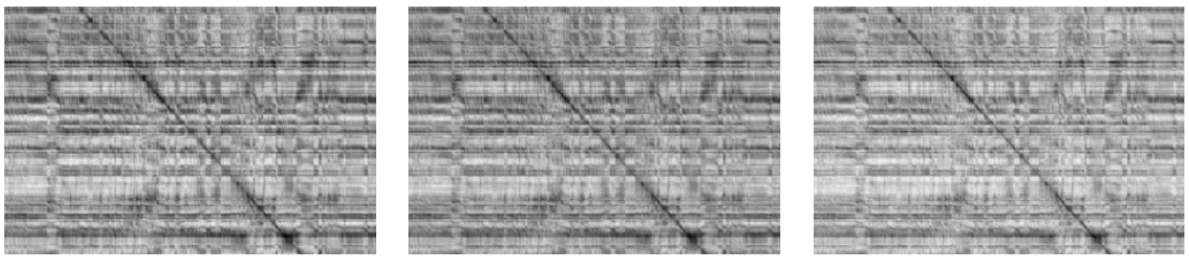
1.4.3 基于Norland数据集的单目图像视觉识别与定位

我们在Norland数据集上进行了测试，以确定多少个维度最适合耗时和准确性。我们在scikit-learn中使用增量PCA进行大量的图像匹配。IPCA是必不可少的高维数据分析方法之一。IPCA通过线性变换将高维数据转化为低维数据。AlexNet不同层的维度如表1所示。很容易理解，我们保留的维度越多，获得的信息就越多，但也很耗时。所以首要任务是确定我们为每个向量保留多少维度。

图 3



(a) Matching image of 5 dimensions features (b) Matching image of 10 dimensions features (c) Matching image of 20 dimensions features



(d) Matching image of 33 dimensions features (e) Matching image of 51 dimensions features (f) Matching image of 99 dimensions features

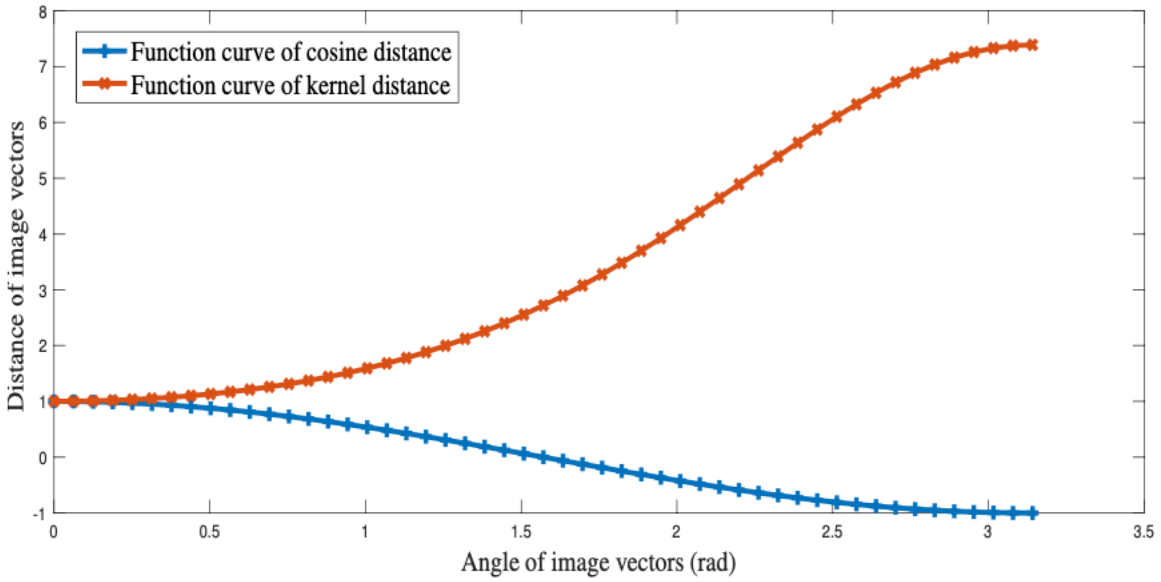


图 4

表2列出了参数n_components和主信息比之间的关系。一般来说，在影响精度的情况下，我们最好保持至少90%的主信息比。我们还比较了不同维度之间的匹配结果。匹配结果如图 3 所示：在20个维度以下无法检测出最佳匹配线。33 个维度既清晰又节省计算量。简而言之，我们选择 33 个维度的向量作为图像描述符。

我们的任务是精确地找到最佳匹配线。我们必须使用数学变换来使这条线更清晰。我们选择的是内核法，包括对匹配矩阵的元素进行反演和指数化。选择这种方法的原因如下：

1）2幅图像之间的余弦距离不能代表相似度和匹配矩阵元素之间的正比例。

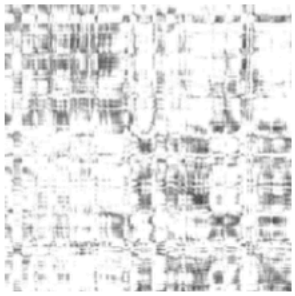
2）内核法会扩大假阴性与真阳性之间的距离。

图4是由余弦距离（如式公式（1）所示）和内核法距离（如式（2）所示）计算出的函数曲线比较。蓝线代表两个图像向量的余弦距离，棕线代表核法距离。棕线代表的是内核法的距离。我们可以看到，内核法可以增强完全不同和相似的地方之间的差异。最佳匹配线的颜色显示为黑色，不同的地方显示为白色，如图5所示。

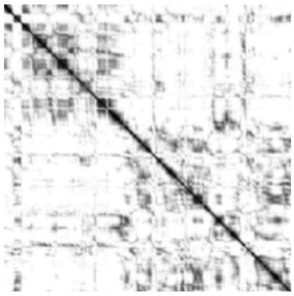
此外，通过核方法将匹配矩阵归一化，减少了大部分在线图像匹配数据集混乱造成的歧义。将匹配矩阵保存为灰度图像，用于后续的处理，包括形态变换和二值化。

更重要的是，我们对匹配矩阵进行了归一化处理，其范围为0～255，公式为（3）。经过内核法处理后，效果明显。这对形态学处理和可视化有很大的帮助。

我们对Norland数据集的春冬两季进行了内核方法测试。有3000张春天的图像和3000张冬天的图像是在同一个地方拍摄的。此外，两个图像序列的开始是相同的图像。因此，一条线出现在对角线上是最佳匹配序列。匹配结果如图5所示。 我们将在线图像与记录的数据集图像逐一通过余弦距离cos < fi , fj >进行匹配。然而，图5(a)所示的匹配图像出现了可怕的匹配和完美匹配之间的混乱。然而，通过公式（2）的核方法和公式（3）的归一化方法，对角线变得明显。匹配图像如图5(b)所示。最后，将匹配矩阵保存为灰度图像，通过适当的阈值将其转换为二进制图像。



(a) Matching matrix of cosine distance



(b) matching matrix after kernel method

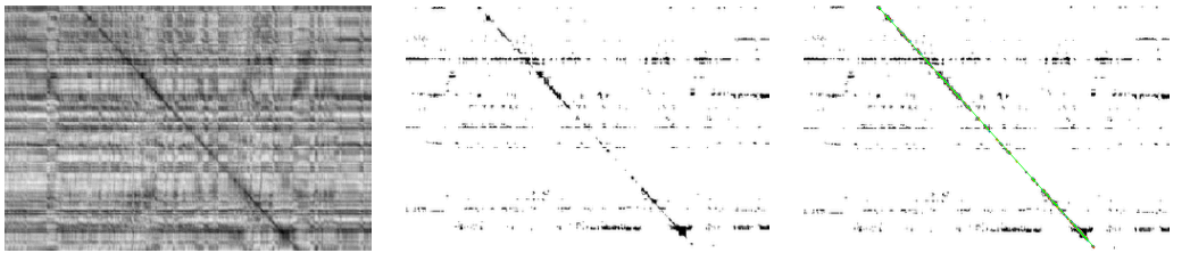
图 5

$$\cos < \mathbf{f}_i, \mathbf{f}_j >= \frac{\sum_{i=1}^{33} a_i b_i}{\sum_{j=1}^{33} a_j^2 \sum_{k=1}^{33} b_k^2} \tag{1}$$

$$\hat{\mathbf{m}}_{ij} = e^{1-\cos \mathbf{m}_{ij}} \tag{2}$$

$$\mathbf{M}_{ij} = \frac{255\left(\mathbf{M}_{ij} - \mathbf{M}_{min}\right)}{\mathbf{M}_{max} - \mathbf{M}_{min}} \tag{3}$$

我们的实验旨在展示我们的方法在减少特征和图像处理的情况下的能力。我们的方法能够(i)在跨季节的场景中进行定位，忽略动态物体、不同天气和季节变化。(ii)节省时间和计算消耗。我们在SeqSLAM[20]中使用的公开的Norland数据集上进行评估。灰色图像每秒钟采集1帧，尺寸已被裁剪成64x32。如果我们的方法在这种不清晰和微小的图像中仍然有效，那么它可以节省大量的时间和计算消耗。匹配图像的例子如图3所示。



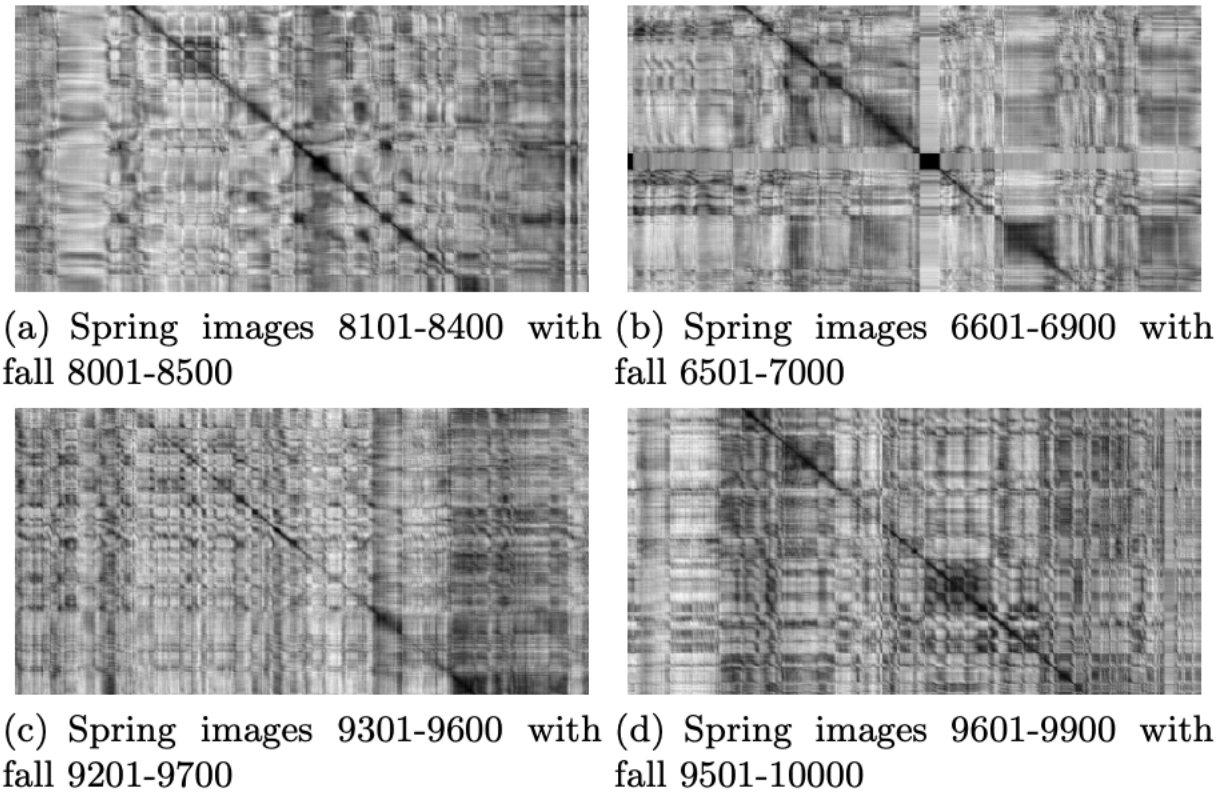
(a) Matching matrix after kernel method and normalization

(b) Binarization image

(c) Best matching line in matching image

图 6

我们可以看到，在图5(b)中，最佳匹配线变得明显。我们的任务是找到它的数学模型，在数据集中找到相应的指数。我们决定使用经典的RANSAC算法。在图6中，我们选择了秋季季节训练的300张图片序列作为地图，在春季季节进行在线定位。我们可以看到，从AlexNet的Conv3中提取的特征并没有影响匹配结果。相反，减少了背景信息的影响，如图6（a）所示。图6(b)是匹配图像的二值化结果。你可以看到，大部分干扰信息已经被擦掉了。在机器人定位过程中，约束分心器的能力有更重要的作用。在图6(c)中，我们可以看到，绿色的线条正是这一时期的最佳匹配。匹配矩阵中当前图像的最佳匹配特征为f_{km+b}。那么当前图像在视觉图中的最佳匹配图像为l_{km+b}。在图8中，我们绘制了3条线来评估我们方法的误差。蓝线代表地面真实的指数。红线是指与我们的方法匹配的图像的指数。黄色的是地面真值和匹配图像之间的指数误差。在x坐标轴[1872，2026]范围内的搜索指数无法更新。



(a) Spring images 8101-8400 with fall 8001-8500

(b) Spring images 6601-6900 with fall 6501-7000

(c) Spring images 9301-9600 with fall 9201-9700

(d) Spring images 9601-9900 with fall 9501-10000

图 7

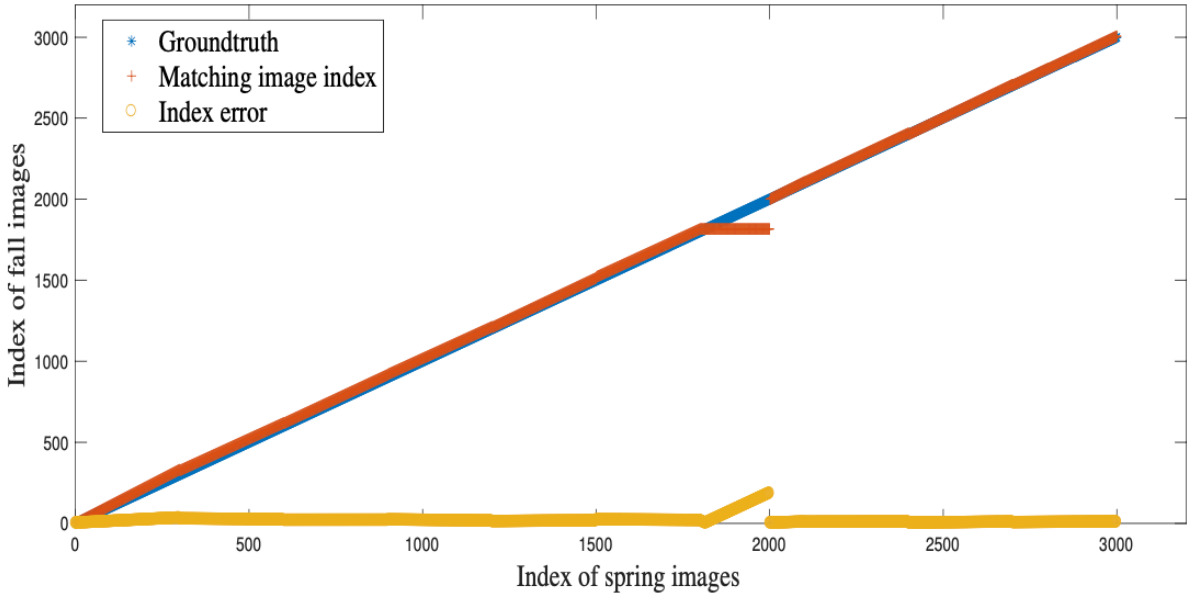
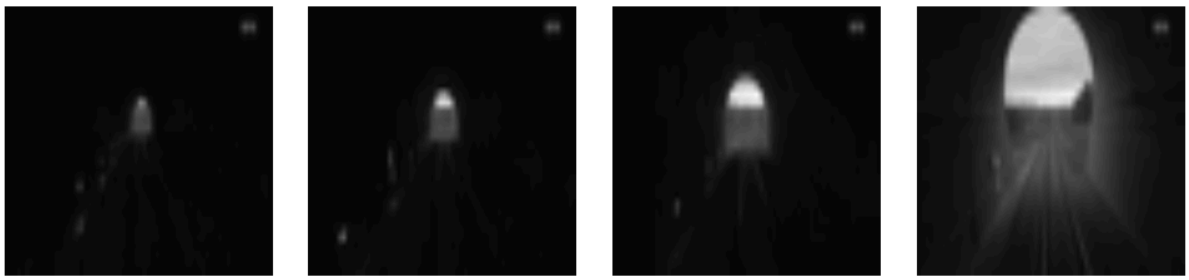


图 8

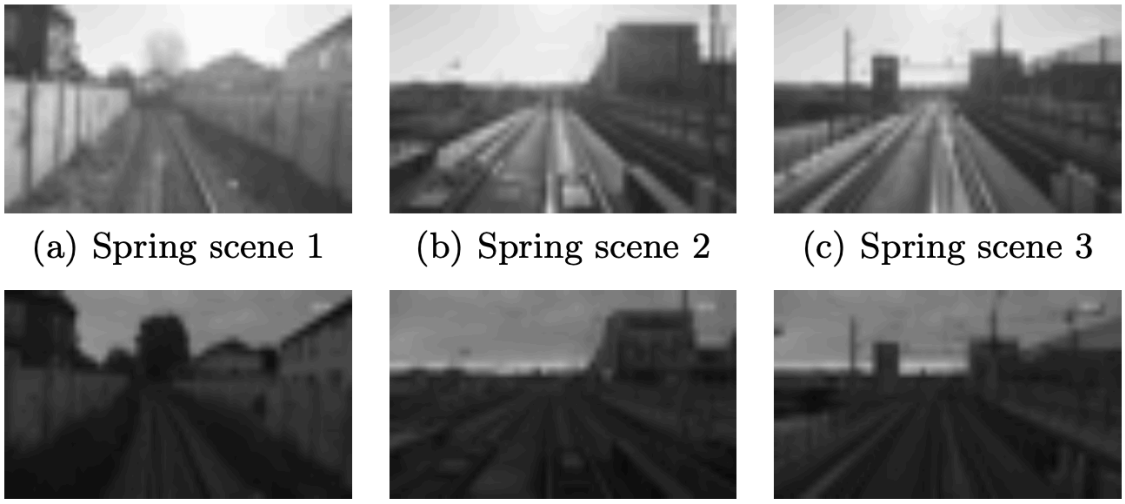


(a) Images-02204 in spring datasets (b) Images-02205 in spring datasets (c) Images-02206 in spring datasets (d) Images-02207 in spring datasets

图 9

我们的论文提出了一种新颖且耗时的算法，在跨季节的动态环境中对机器人进行定位。这是一个快速定位系统。我们从AlexNet的Conv3中提取特征，它在机器人定位领域确实优于手工制作的特征。通过IPCA减少尺寸是一个新的尝试。AlexNet的每一层都会在不同的领域发展出优势。事实证明，Conv3是机器人本地化的最佳选择。我们通过内核法距离将在线图像的向量与数据集向量逐一进行比较。这个过程扩大了相似和完全不同的地方之间的差异。此外，对灰度匹配图像进行图像处理，包括通过适当的阈值转换为二进制图像，将复杂的数据关联图转化为简单的图像处理。在序列匹配方面，我们采用经典的RANSAC算法来寻找最佳匹配线。我们的

实验结果表明，减少维度是加快计算速度和减少混乱匹配的好主意。而且我们的算法对季节变换、动态环境、天气变化等都有很好的鲁棒性。一些匹配图像的例子如图10所示。



(a) Spring scene 1 (b) Spring scene 2 (c) Spring scene 3

(d) Fall scene 1 (e) Fall scene 2 (f) Fall scene 3

图 10

我们系统的局限性来自于图像采集设备。在完全黑暗的环境中，图像很难表达。其实在图8中，1872到2026的图像没有匹配线，所以我们根本无法检测到匹配线。图9是暗图像的例子。图9为暗部图像的例子，匹配图像为黑色块状。我们将考虑增加激光的辅助。此外，我们还将研究特征尺寸与定位精度之间的具体关系，找出最合适的CNNs特征尺寸，以保证精度和运行速度。