Pose Pooling Kernels for Sub-category Recognition

**摘要**

在训练数据有限的情况下，基于超类别界标对姿势进行归一化的能力可以显着改善单个类别的模型。 先前的方法已经考虑将体积模型或可变形模型用于面部和某些类的铰接对象。我们考虑对关注类别施加较少代表性假设的方法，并利用考虑到针对特定姿势关键点配置训练的检测器的响应整体的当代检测方案。 我们使用显式翘曲和隐式池作为机制开发基于姿势的姿势归一化表示。 我们的方法定义了适合于最近邻居或基于核的学习方法的姿势归一化相似度或核函数。

**1.简介**

对于现代计算机视觉系统，细粒度类别的识别是一项重大挑战。 这样的类别可以通过相对局部的特征来区分，这些特征可能很难从常规的1-vs.all学习框架中的有限数量的训练数据中学习。 当一组相关类别共享某种结构时，可以从跨多个类别汇总的数据中学习姿势估计器。 一般而言，基于超类别界标或姿态检测器对姿态进行归一化的能力可以显着提高有限数量的训练数据对各个类别的识别。

姿势归一化的方法早已用于人脸识别[9，18]。 对于凸对象，可以将姿势建模为刚性运动，还可以选择进行非刚性变形。当考虑到更一般的铰接对象类别时，姿势估计的问题变得更加复杂。 最近，基于地标模板或基于“ poselet”的姿势估计已成为人们越来越感兴趣的话题[7、5、4]。 在我们先前的工作中[12]，我们利用这种模型来构建可对姿态进行操作的姿势归一化描述在有关节的物体上。 但是，该模型需要实例化3-D体积基元以形成表示形式，这种表示形式的获取成本很高，而且手动获取可能会出现问题（请参见图1）。

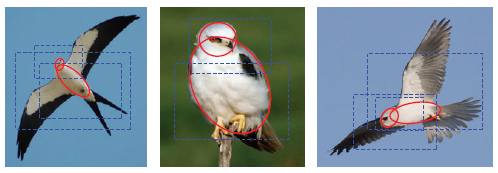


图1.头部/身体体积表示的局限性。诸如[12]中所示的体积表示（红色椭圆形）将不足以确定栖息的鸟类匹配飞行中的两只鸟类中的哪只。 机翼和机尾（颜色和形状）几乎都具有可辨别的外观信息，并且可以用一个姿势集（蓝色虚线框）进行建模。 你能分辨出那只鸟匹配吗？

在本文中，我们还解决了子类别识别的几何归一化问题，但主张采用2-D而不是3-D表示。 我们假设一个检测模型为[7，5，4]样式，该模型在给定图像上产生一组姿势模型激活，并探讨如何最好在两幅图像之间比较这类检测到的特征的问题 。我们开发了相似性函数，这些函数将体态激活“堆栈”作为输入，适用于最近邻居分类器或SVM内核。

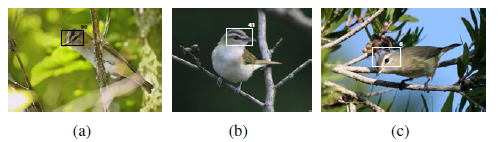


图2.比较Poselet的外观。 对于使用区分性分类器（或最近邻）的子类别识别，我们需要一种机制来比较姿势集。 三个不同的姿势可能实际上以不同的姿势覆盖了相同的基础部分。 因此，我们需要一种基于这些姿势来比较外观的方法。 你能分辨出哪两只鸟是一样的吗？

图2展示了我们相似性函数背后的关键思想，其中展示了在不同的鸟类实例上射击的三个不同的小体。右边的两个图像描绘了同一子类别的实例； 例如，使用空间金字塔匹配的内核或包或单词进行整个图像（或整个鸟类）的比较可能会错过两只鸟类外观上的明显对应关系。但是，通过利用示例中两个姿势实际上重叠相同部分的知识，我们可以定义一个比较函数，该函数显式比较两个图像中相应姿势区域上形成的描述符。

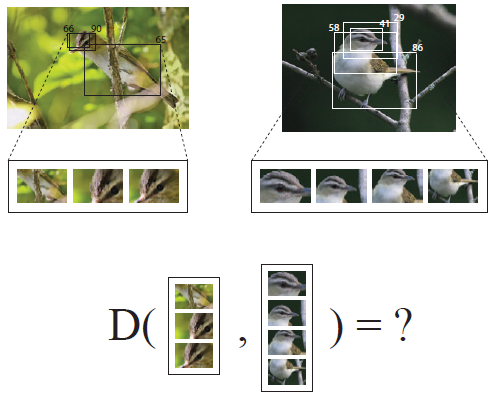


图3. Poselet集相似度的图像相似度。 我们建议通过定义一系列姿势集相似性度量来测量图像相似性。与其考虑全局图像内的图像统计数据，我们提倡使用姿势集作为将图像块内的对象外观与训练数据中语义相似的部分相联系的一种手段。 这有效地提供了高度的姿势不变性。

我们定义并比较了一系列的姿势集相似性度量或核。一个直观的想法是将贪婪匹配内核与具有里程碑意义的对应关系一起使用具有显式几何变形的贪婪匹配内核，从而构造一个贪婪地估计最小成本对应关系的匹配内核。这种方法很优雅，但是在大多数情况下在计算上都是棘手的。 然后，我们考虑形成固定长度向量的表示形式：一个变体尝试使用翘曲函数在每个示例中对表示形式进行规范化，而一个更简单的模型则将描述符存储在相应的姿波上。 我们的池化方案根据每个Poslet展示的重叠程度在Poslet之间建立对应关系：从概念上讲，目标是为实际上覆盖相同零件的Poselet的描述符进行池化。

我们在最近引入的CUB鸟类数据集上评估了我们的方法，比较了给定嘈杂检测的各种描述符的识别性能。 总体而言，与未针对姿势进行规范化的基线方法相比，我们从提议的池体系结构中发现了显着的提升。 我们的结果表明，即使对于不接受可靠体积描述的类，有效的姿势归一化也是可能的。 尽管我们的实验仅限于鸟类领域，但我们希望我们的姿势库核可用于其他各种识别任务，这些任务中姿势差异很大，但每个类别的训练数据有限。

**2.背景**

以前关于下级分类的工作包括学习判别式图像特征的方法。 先前已考虑过的此类域包括：花的从属类别（Nilsback和Zisserman [25，26]，介绍了17和102类牛津花的数据集），六个基本对象类别（例如，Grand vs Upright Pianos（Hillel et al。[2]）中的每一个都有两个子类别。 以及石蝇幼虫的下属类别，它们具有巨大的视觉相似性（Mart´ınez-Mu〜noz等[23]）。

大量文献试图利用类别之间的相似性来提高识别性能。利用类分类法或层次结构的方法从构造潜在主题层次结构[3]到共享外观[16，27]，分类器[1]或视觉部分[28]到构建有效的分类树[17，22]以及其他参考 这里要提的很多。每种这样的方法都为有效解决基本级别分类提供了见识或进步。但是，这些无监督的方法不能轻易地应用于区分密切相关的从属类别的问题，这些从属类别根据定义具有共同的部分集合，但是可能同时具有细微和剧烈的外观变化。

一些作者研究了基于属性的识别，这些属性与子类别识别的一般问题高度相关，例如，参见[10、11、19、20、33]。 这些技术通常从带有属性标签的训练数据中学习判别模型，然后应用所学习的模型来估计测试图像中存在的适当视觉属性。 尽管基于属性的模型适用于解决单样本学习问题（之前在[13、14、15、24]等中曾考虑过），但它们通常专注于相对粗粒度的属性。我们的重点是适合于下级分类所需的精细区分的表示形式。

布兰森等人的工作[8]提出通过将计算与对人类对象的属性查询进行交织来提高识别精度。 他们的方法在一个大型200类鸟类数据集中评估识别度[34]，这也是我们实验的主题。

如[7，5]中所述，我们的方法基于poselet框架，另请参见[4]的相关技术。 我们在此框架中探索了子类别外观描述符的姿势归一化的思想，该主题先前在[12]中已考虑。探索的范例是使用[3,9]建立的针对人脸识别的工作线，使用3-D原语进行体位归一化。 但是，与[12]相反，我们探索了一种具有较少表示假设的方法：特别是我们的方法不采用体积表示，因此适用于不严格接受这种模型的对象类。 此外，更重要的是，我们的方法不需要[3]中的3D姿势注释，就像[12]中的方法一样。 姿势动物社区的最新工作已经考虑了活动识别和属性描述的任务[6]； 这项工作计算了由姿势小子检测器激活组成的特征向量。相比之下，我们的方法（和[12]的方法）在局部姿势小波检测上形成描述符； 本文的贡献在于定义和分析了各种2-D方案，用于比较基于姿态的描述符集，以使与对象的相同基础部分或区域相对应的姿态对齐，从而使对应的描述符可以对齐。有效地比较。

**3.姿势归一化内核**

给定学习的姿势的整体，姿势的检测方法（上面已概述）可以为每个图像推断一组检测。 我们的目标是使用这些检测来计算可有效区分例如单个物种的子类别级别描述符。 特别地，我们想探索用于比较两个检测到的特定姿势模型的实例的姿势归一化外观的方案。 我们在每次姿势运动激活时都计算描述符，并在以下小节中考虑用于比较这些描述符集的各种策略。

为了将判别式分类器用于子类别识别，因此，我们需要一种机制来比较两组姿势检测。 姿势检测过程可提供零件位置的估计值； 我们的推测是，与仅使用整个图像而不进行任何姿势归一化相比，比较与相同物理部分（或部分集合）相对应的图像描述符将提高分类性能。 通常，子类别识别取决于某些部分的细微外观变化：两个不同的姿势可能实际上只是以不同的姿势或视图覆盖了相同的基础部分，因此希望有一个相似度函数可以适当地关联来自 比较响应集时的各种姿势检测。 我们在下面考虑解决此问题的各种方法，包括在比较描述符之前通过几何扭曲来计算体形到体形归一化的方案，以及在对应的（语义上相似的）体形上合并描述符的方案。

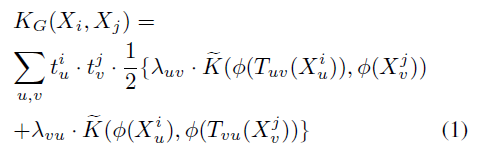
为了将最近的邻居分类器和基于核的分类器直接应用于我们的子类别识别问题，我们基于检测到的姿势集定义了核函数。 可以在例如基于SVM或高斯过程的分类器或回归方案中使用这些功能。

**3.1. Preliminaries**

每个图像Xi具有一组具有相应的激活得分的微调激活窗口，其中N是poselets的数量。 假设我们从每个姿势的u激活窗口中提取d维图像描述符，例如单词袋SIFT或空间金字塔直方图。 然后，每个图像可以表示为。 同样，在每对姿势球u和v之间，我们计算从姿势球u到姿势球v的变换函数Tuv以及该变换的置信度。

仿射变换函数Tuv是基于两个姿势的关键点位置计算的。 如果两个姿势组的共同关键点少于三个，则两组关键点之间将不会进行适当的仿射变换。 在那种情况下，Tuv设置为空，置信度设置为零。 否则，我们从仿形体u的关键点集到仿形体v的关键点集计算仿射变换Tuv，并基于重叠的关键点数设置置信度得分，即，其中K为 共同关键点的数量，Ku是poselet u的关键点的数量。

理想情况下，我们首先根据[32]的精神考虑匹配核，该核将通过将一个图像中的每个姿势检测转换为第二个图像中的另一个姿势检测来比较两组姿势激活，然后比较相应的图像描述符。 贪婪扭曲匹配内核的定义如下



其中是对齐的小体之间的基本内核，是将激活窗口从小体u弯曲到小体v后的图像描述符； 取两个翘曲方向的平均值可使内核对称。权重是基于重叠点数的变换的置信度。

如下面更详细地描述的，用于对准的小体的合适的核函数可以是简单的线性核或非线性核，例如卡方距离，其是在每个检测到的小体位置提取的梯度直方图描述符之间计算的。

使用该内核，每对图像之间的相似度只是姿势归一化图像描述符之间的相似度的加权和。 该核函数可能是有效的，但是当检测到的姿势数量很大时，会遭受很高的计算成本。 这需要O（n2N2）时间，其中n是图像数，N是训练后的姿势数。 因此，在涉及大型数据集的情况下，此方法可能无法很好地扩展。 因此，在以下各节中，我们将考虑中间固定长度表示形式，这些表示形式通常采用扭曲或更直接地合并以对齐相应的姿波。

**3.2扭曲的功能内核**

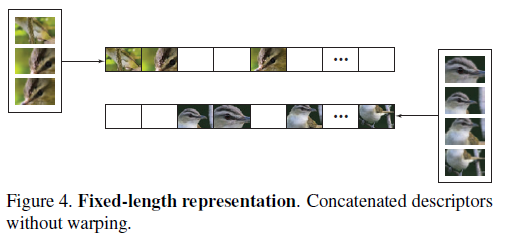
为了克服天真匹配内核的二次复杂性，该天真匹配内核显式比较检测集，我们考虑使用固定长度表示形式来捕获对象的姿势图集。 由于这定义了向量空间，因此可以将其直接用作特征向量，例如卡方或径向基函数（RBF）内核。

最直接的表示方式只是将每个姿势的图像描述符连接到一个长的固定长度特征向量。 这简单地表示了图像在不同姿势下的外观，并用作基线方法。 但是，如果没有几何归一化，除非可用训练数据覆盖所有类别的所有可能的姿态激活，否则该特征向量的性能将很差。

按照上面的表示法，简单的固定长度表示为



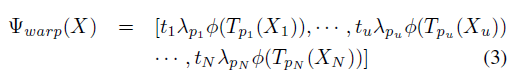
其中是微调u的激活窗口的图像描述符，而tu是激活分数。 该特征向量的长度为Nd，其中d为图像描述符的维数。 图4说明了此方法。



这种特征表示的一个重要问题是特征向量稀疏，因为在典型图像中仅检测到少量的姿态（在下面描述的数据中，我们的实验中为10个）。 同样，该表示可能是多余的，因为不同的姿势通常是重叠的，因此它们以不同的姿势和视图描述相同的对象区域。 为了克服这个问题，我们考虑了一些方法来对该表示进行姿态归一化。

我们的第一种方法遵循上述考虑的定长表示的精神，并在定长表示内显式扭曲姿势的外观以填充尚未在图像上触发的姿势。 有效地，这将以固定长度表示形式填充空白特征块。例如，假设19号姿势和23号姿势都以略微不同的方向捕获了鸟头的左侧。 对于一张显示鸟头左侧的图像，它可能只是发射了19号姿势，而在另一幅图像中，23号姿势就将​​发射。 这两个姿势子都代表鸟类的相同部分（头部的左侧），如果可以在特征向量表示中捕获这种对应关系，它将改善分类。

因此，对于式中的每个。 对于尚未检测到但存在另一个与之足够相似的检测到的姿势的图2，我们使用发射姿势的图像描述符并将其扭曲为未姿势姿势。 通过这种方法，特征表示为



其中pu是应该转换为非发射姿势u的最相似发射姿势的索引。 如果该姿势已被触发，则它会坚持公式。 如图2所示，如果没有合适的发射姿势，则未发射姿势的相应特征将设置为零。 我们将变换后的残留误差用作两个体形相似度的度量。 图5说明了此方法。

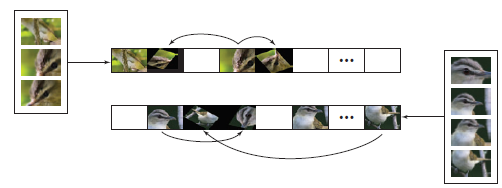


图5扭曲的功能内核 带变形的串联描述符。

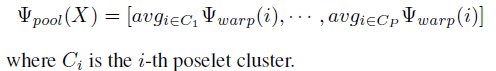
**3.3合并功能内核**

固定长度变形内核背后的直觉是要采用姿势归一化的方式来比较在不同部分具有对应关系的图像。 该模型的进一步扩展是将代表相同基础部分的物态分组或合并为部分簇。

通过设计，学习过的姿势集会表现出冗余性：几个姿势集将代表相同的零件，或者相对于相机而言，其配置略有不同。为了识别，比较表示时最好将它们组合在一起。 因此，我们在基本表示的基础上考虑一个合并阶段，该阶段将在姿势集上计算出的描述符组合在一起，这些姿势被标识为是对应的。 此策略对于加性内核（例如在局部特征上形成的单词袋表示）特别有效，但在非加性表示上也可以发挥一定作用。

我们考虑了两个将姿势组分组的标准，每个标准都包含代表对象同一部分的姿势。 可以根据提供的零件注释以完全受监管的方式处理此问题； 但是，我们选择考虑以数据驱动方式发现集群的无监督方法。

如图6所示，我们的合并方案形成了一个聚类特征向量，其长度等于聚类数乘以姿势描述符的长度。 对于每个聚类，描述符在分配给该聚类的姿势集上合并，从而为聚类生成单个描述符。 最终的聚类特征向量是聚类描述符的串联，如以下等式所示：



理想情况下，每个姿势簇都应对应于一个连贯的零件或零件组，并且每个组中的所有姿势都彼此相似。 使用这种聚类方案，输出池图像描述符在描述不同部分的图像特征方面更具代表性。

我们使用贪婪聚类方案计算体形聚类，该算法首先在学习的体形上形成图，并计算边缘距离以反映逆体形对应的度量。 我们对边缘距离使用了两种不同的度量：

1.扭曲距离-使用对应于两个姿势的关键点之间的仿射变换的残差。

2.关键点距离—根据两个体态共有的关键点数，如上定义的1 /。

导致明显的聚类结果。 下面我们根据分类性能比较两个合并结果。我们随机选择姿态集作为候选聚类中心，并根据上述两个条件之一将足够数量的邻居分组在一起。 我们重复进行，直到所有姿势都被迭代地分配给聚类中心。 具体来说，聚类算法首先随机选择一个姿态集作为聚类中心，然后将其余的姿态集进行分组，这些其余的姿态集的距离都在设置的阈值之内。 然后，迭代地选择另一个未选择的体形作为新的聚类中心，并重复该过程，直到没有好的聚类。 该方法的优点是不需要先验地了解集群的数量。 其他聚类方案可能证明优于该贪婪方法，并且将成为未来工作中要考虑的领域。 上述方法在实践中效果很好。

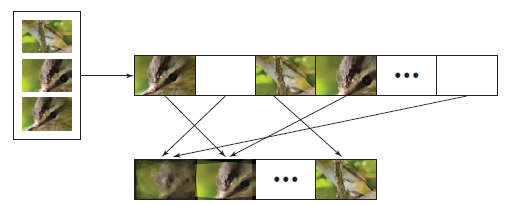


图6.姿势池化内核。 相应的姿势组被分组。