计算机视觉中的稀疏与低秩

摘要

Abstract

目录

第一章 绪论

1.1 研究背景及意义

1.2 研究现状概述

1.3 本文的研究内容和创新点

1.4 论文结构安排

第二章 稀疏与低秩问题概述

第三章 稀疏表示在BoW模型中的应用

3.1 引言

讲清楚为什么要改进BoW模型

近些年来，在计算机视觉领域的诸多应用之中，例如图像或视频的分类或检索、目标的检测和识别、物体的分割等等，BoW（Bag-of-Words，或称做词袋）模型作为一种表征图像或图像区域的方法得到了极为广泛的应用。它的出现主要是为了解决一幅图像或者图像中的某一区域中描述子的数量不确定或者是过于庞大的现象。通常情况下，为了获取更细致的表述，人们常常采用局部描述子。但是，由于描述子通常选取在具有显著纹理特征的位置，例如角点，而一幅图像中这些位置的个数常常是不固定的，因此描述子的数量也是不定的。另外，还有一种与选取描述点不同的叫做密集采样（Dense Sampling）的方式，它在具有固定间隔的位置进行描述，如6个或者8个像素。这种方式尽管能够更加全面的描述图像，但是也带来了描述点数量过多的问题。为了解决以上两种问题，BoW模型采用了量化的方法：给定一组基（即模型中的words，或称做visual words），将每个描述子量化为这组基中的某一个，然后在图像或区域中进行频率统计，最终形成一个具有固定长度的直方图。

在BoW模型中，量化使用的基通常是在整个数据集上由聚类算法产生的，其中比较典型的是K-Means算法。这种聚类算法能够保证在欧式空间中基于整个数据集的距离最小，但是这种无监督的方式是不包含有任何类信息的，因此，得到的基对不同的类也不具备区分性。并且，直方图的形成依赖于给定的基，基的好坏将直接影响量化的性能。不幸的是，在这组基的形成过程中，噪声没有被滤除，而是掺杂了进来，这进而会影响量化的结果。

针对这些问题，本章提出了一种依靠稀疏表达寻找主要基的方法。在该方法中，通过利用描述子的类信息，将去除那些不重要或者充满噪声的基，并建立一组更有区分性的基。描述子虽然处于语义描述的最底层，但它同样带有类信息。例如，用于刻画草地或绿树的描述子是很低的概率出现在天空或是大海中的。因此，可以利用这种信息去除那些在聚类过程引用的噪声，并且最终形成的基将更加具有区分性。比如说，在描述楼房这一语义的时候，具有垂直、水平、或者90°拐角纹理体征的描述子比其他的类型更加的有效。在生物学中有着实际意义的稀疏表达，在某种意义上具有针对基的选择功能。当一幅图像在一组基上进行稀疏的表示时，那些被激活的基实际上就是组成该图像的最基本的元素，稀疏表示也因此而完成了它的选择功能。基于这种特性，我们可以使用这种选择功能为分属于同一类的描述子去寻找其共通的基，同时这组基也具备区分该类和其他类的特性。

我们将对本章剩余部分做如下的结构安排：在3.2节，我们将着重介绍BoW模型；然后在3.3节详细描述如何使用系数表示获取BoW模型中重要的基以及整个算法的流程；3.4节讲述了为了验证算法有效性而在三个数据集上的实验、结果以及分析；最后我们在3.5节给出本章的总结。

3.2 BoW (Bag-of-Words)模型

讲清楚什么是BOW模型

3.2.1 由自然语言处理到计算机视觉

BoW模型最早是被用于自然语言处理中【Harris, Zellig (1954). "Distributional Structure". Word 10 (2/3): 146–62.】。在这个模型中，一段文本（比如一个句子或者一个文档）被用一组无需的词来表示，而不需要考虑语法方面的因素。然后，统计每一个词的发生频率用作该段文本的特征表示。例如，这里有两个简单的文本文档：

1) John likes to watch movies. Mary likes too.

2) John also likes to watch football games.

接下来我们统计文档中的词以及出现的次数：

{"John": 1, "likes": 2, "to": 3, "watch": 4, "movies": 5, "also": 6, "football": 7, "games": 8, "Mary": 9, "too": 10}，

这样就形成了一个由10个不同的词组成的字典。然后依照词典中词的索引，每一个文档都可以表示为一个10维的向量：

[1, 1, 1, 1, 1, 0, 0, 0, 1, 1],

[1, 1, 1, 1, 0, 1, 1, 1, 0, 0].

向量中的每一个元素表示了词典中的词在文档中出现的次数。这种方法虽然忽略了单词的顺序以及语法的信息，但是依然在一定程度上代表了原始的文档。

近些年来，BoW模型被移植到了计算机视觉领域，其最早可追溯至【L. Fei-Fei and P. Perona (2005). "A Bayesian Hierarchical Model for Learning Natural Scene Categories". Proc. of IEEE Computer Vision and Pattern Recognition. pp. 524–531.】。 在这篇文章中，作者使用了相同的规则，将一幅图像在一组训练好的码书（Codebook）上面进行表示，得到了一个具有固定维度的直方图，如图3.1左侧图所示。与自然语言处理中相似的是，“文档”被看做了“图像”，“字典”被看作了“码书”，字典中的“词”替换为“视觉词”（Visual Words），因此在计算机领域中BoW也常被称作Bag-of-Visual-Words（BoVW）。

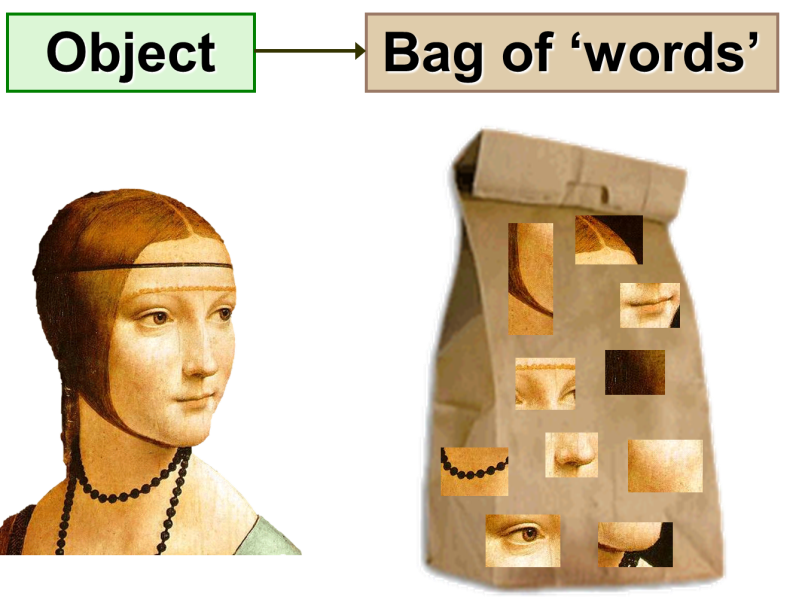


图3.1 BoW模型示意图。在BoW模型中，一幅图像被若干个与该图像相关的视觉词汇所表示，如左图所示。图片来自【L. Fei-Fei, R. Fergus, and A. Torralba. "Recognizing and Learning Object Categories, CVPR 2007 short course"】。实际操作中，每一幅图像则被表述成为一个直方图向量，如右图所示。

Bow模型在计算机视觉中取得了非常广泛的应用，究其原因，大致可分为以下两点：1）简单性，无论图像或区域的具有怎样的面积，它都能使用一个具有固定长度的直方图进行表示，这使得该图像或区域的语义表述变得极为简单，同时简化了不同图像或区域之间的相互比较和区分；2）有效性，具备相同内容的图像或区域一般具有相似的直方图，反之亦然。例如图3.1右侧图，上中两幅图同为绿树，因此有着相似的直方图表述，但这与沙漠形成的直方图却是截然不同的。这种特性使得模型形成的直方图具有一定的代表性，因而能够被用于区别不同语义概念。

3.2.1 计算机视觉中的BoW模型

毫无疑问，BoW模型凭借其简单却有效的功能在计算机视觉中起着重要的作用。通常意义上讲，该模型可以分为以下两个部分：1）码书构建；2）描述子量化；3）汲取。下面，我们就这三个方面进行详细的描述。

3.2.1.1 码书构建

模型所需要的码书是在整个数据库上面进行构建的。如图3.2所示，首先，需要从数据库中的图像上抽取出全部的描述子，然后再使用聚类算法得到需要的码书。

在抽取描述子之前，需要先确定关键点的位置。这里通常有两种选取方式，一种是寻找在局部空间以及尺度上变化剧烈的像素点，正如SIFT（Scale-invariant feature transform）特征或者它的改进版本SURF（Speeded Up Robust Features）特征所采用的关键点检测方式。这种选取方式一般位于纹理较丰富的区域，比较符合人类的视觉感应系统，因为人总是对变化剧烈的区域比较敏感。但是，对于那些仅包含平坦区域的语义概念，如天空、大海等，通过这种方式选择出来的关键点在数量上非常的少，进而可能影响到对图像的表示性能。密集采样就是针对这一问题而提出的改进方式，它通过包含更多的关键点来获得更加细致的描述。图3.2中所示意的关键点的选择就是采用的密集采样方法。局部描述子的类型则是多种多样的，例如前面提到的SIFT和SURF，还有MSER（Maximally Stable Extremal Regions）和HOG（Histogram of oriented gradients），以及由此扩展出来的变种（如Color-SIFT）等等。但就性能而言，SIFT是最好的，因为描述子的性能问题不是本文讨论的重点，因此不再多加描述。

得到描述子之后，需要进一步将其聚类形成所需要的码书。K-Means是一种常常被采用的聚类算法，当然，其他的一些快速算法，例如Hierarchical K-Means采用了分级聚类的策略，Approximate K-Means[Muja, M. and Lowe, D. Fast approximate nearest neighbors with automatic algorithm configuration Proceedings of the International Conference on Computer Vision Theory and Applications (2009)]采用了近似算法，在速度上得到了提升，也因此被一些文献所采用。码书的大小，即K-Means算法中的K，是需要提前指定的。在常用数据数据库上，比如Caltech或PASCAL等，这个值通常被设置为1000或者2000左右，当然取值越大，区分性会越好，但同时计算量也会随之而增大。假设抽取出的N个描述子为，K-Means的目标就是寻找包含有K个码字的码书，使得它满足：

这样就满足了在欧式空间中，码书和描述子之间的距离最小。由于描述子来自于整个数据库，数量过于庞大，因此实际操作中需要进行适当的下采样。

3.2.1.2 描述子量化

通过聚类得到码书之后，就可以将描述子分配到各个码字上，这一过程称之为量化。量化方式有很多种，最常用的一种就是最近邻方法，也就是将一个描述子分配到与之距离最近的码字上面，即：

这里，描述子被分配至第个码字上面，并赋予1的权重。这种方式也被称作硬量化（Hard Quantization），就是说描述子被分配到距离最近的码字上，同时给予1的权重。与之相对应的是一种称作软量化的（Soft Quantization）的方法，它不但向最近邻码字上分配，同时还向M近邻的码字上分配，只是分配的权值不再是1，比如对于第m个最近邻码字，它将分配的权值。其目的主要是处理那些离群点，使得形成的直方图对噪声更加具有鲁棒性。

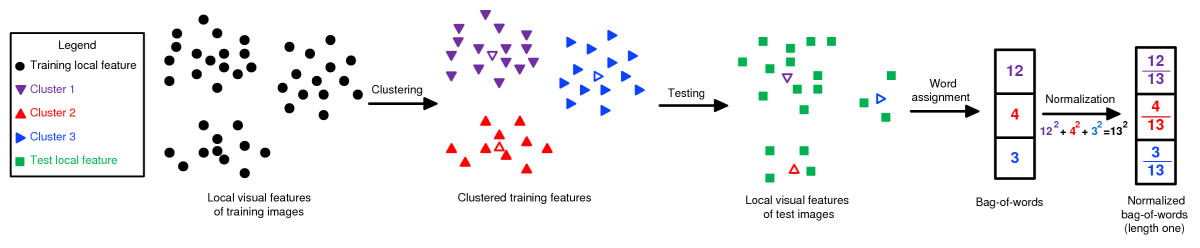


图3.2 BoW模型的三个组成部分：码书构建、描述子量化以及特征汲取。在来自于整个数据库的描述子上面，通过聚类构建出码书。然后一幅新图像的描述子经过量化被分配到每一个码字上面，并给予相应的权值，这样就获得了一组量化后的向量。最后，根据不同的准则，从这些量化向量中汲取出一个特征向量，然后经过归一化后用于表征该幅图像。图中所示的为求和汲取方式，以及范数的归一化方法。

3.2.1.3 特征汲取（Pooling）

当把图像中所有描述子都分配完毕后，每一个描述子都对应一个维的向量，如果将这些向量相加，那么便可以得出一个能反映该幅图像的直方图。如图3.2中所示，从图像中提取出的12个描述子依据最近邻原则，被分配到了三个码字上面，这样就得到了12个量化向量。然后对这些量化后的向量进行加和处理，并使用范数进行归一化，最终得到一个3维的直方图用于表征该图像。这就是汲取（Pooling）的过程，从若干独自的维向量获取一个能代表某一图像的维向量。图中所示的为求和汲取（Sum Pooling）过程，除此之外，还有最小值汲取（Min Pooling）以及最大值汲取（Max Pooling）。我们将上述三种方式表示如下：

这里，表示第个描述子量化后形成的维向量，表示汲取后形成的维向量，它就是可以表示该幅图像的特征向量。文献【】对汲取方式进行了理论分析和实验认证，最终得出最大值汲取是最为有效的汲取方式。

这样，通过码书构建、描述子量化以及特征汲取之后，就形成了一个可以代表原图像的特征向量。

3.2.1.4 空间金字塔匹配（SPM）

为了追求更高的性能，Lazebnik等人在原始BoW模型的基础上加入了空间的信息，提出了一种叫做空间金字塔匹配（Spatial Pyramid Matching，SPM）的方法。其思想主要来源于【】，在文中，作者认为一对点如果真的匹配，那么它们无论是在较粗的描述层面上，还是在较细的层面上都是匹配的。

K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classiﬁcation with sets of image features. In Proc. ICCV, 2005.

Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories

S. Lazebnik, C. Schmid, and J. Ponce, CVPR 2006

如图3.3所示，这种方法首先在三个层次上将一幅图像均匀的划分为，和个子区域，然后在每个子区域中应用BoW模型。最终将每个子区域上形成的直方图串联起来，并依据不同的层次给予不同的权重。这种方式可以看作是全局和局部描述的相结合，左侧的划分是和原始模型相一致的，是在全局范围内的描述，而右侧则是在更加细致的局部区域对图像进行描述。通过这种粗糙与细致的结合，该方法在通用的数据集上取得了比原始模型更好的结果。

图3.3 空间金字塔匹配（SPM）示意图。图像在三个层次上被均匀划分，最终的特征向量是由各个子区域的特征串连而成的。该图来自文献【】。

3.3 依靠稀疏表示寻找重要的码字

讲明原因

众所周知，自然图像中存在有大量的噪声，而在抽取描述子的过程中，这些噪声并没有被移除而是参予了码书的构建过程，因而最终形成的码书在某种意义上讲是存在有噪声的。再者，由于码书构建过程中采用的是无监督聚类方式，虽然简化了过程，但也使得类与类之间的码书不再具有区分性。当区分某一类与其他类时，只有特征向量的其中一些分量在起主要作用，而其他的则显得不是那么重要，甚至有的还会模糊该类的基本特征。比如，在牛津大学创建的花朵库（Oxford Flower Dataset）[http://www.robots.ox.ac.uk/~vgg/data/flowers/]中的黄色百合（Yellow Lily）一类，如图3.4所示，其区别于其他类的主要特征是颜色，而其他的诸如形状、纹理类的特征则不是很明显，甚至还会起到负作用。再如，我们在PASCAL 2007 数据库上的实验。首先将数据库中的从属于飞机（Aeroplane）一类的所有图像抽取描述子，然后在一个事先获取的码书上面进行量化。这里的码书是在整个数据库上使用K-Means算法聚类得到的。仿照BoW模型，我们最终会获得一个能够表示整个类别的直方图。整个过程如图3.5所示。观察这一直方图，我们会发现，直方图向量的一些元素的取值非常小，换言之，在图像表述过程中，这些元素对应的码字是不起任何作用的。因此，我们认为可以通过学习获取一个更加紧凑的表示，在这个表示下，该类的特征能够得到更好的表达，更有利于区分该类与其它类。

图3.4 牛津大学花朵库中的黄色百合类的一些示例图像。

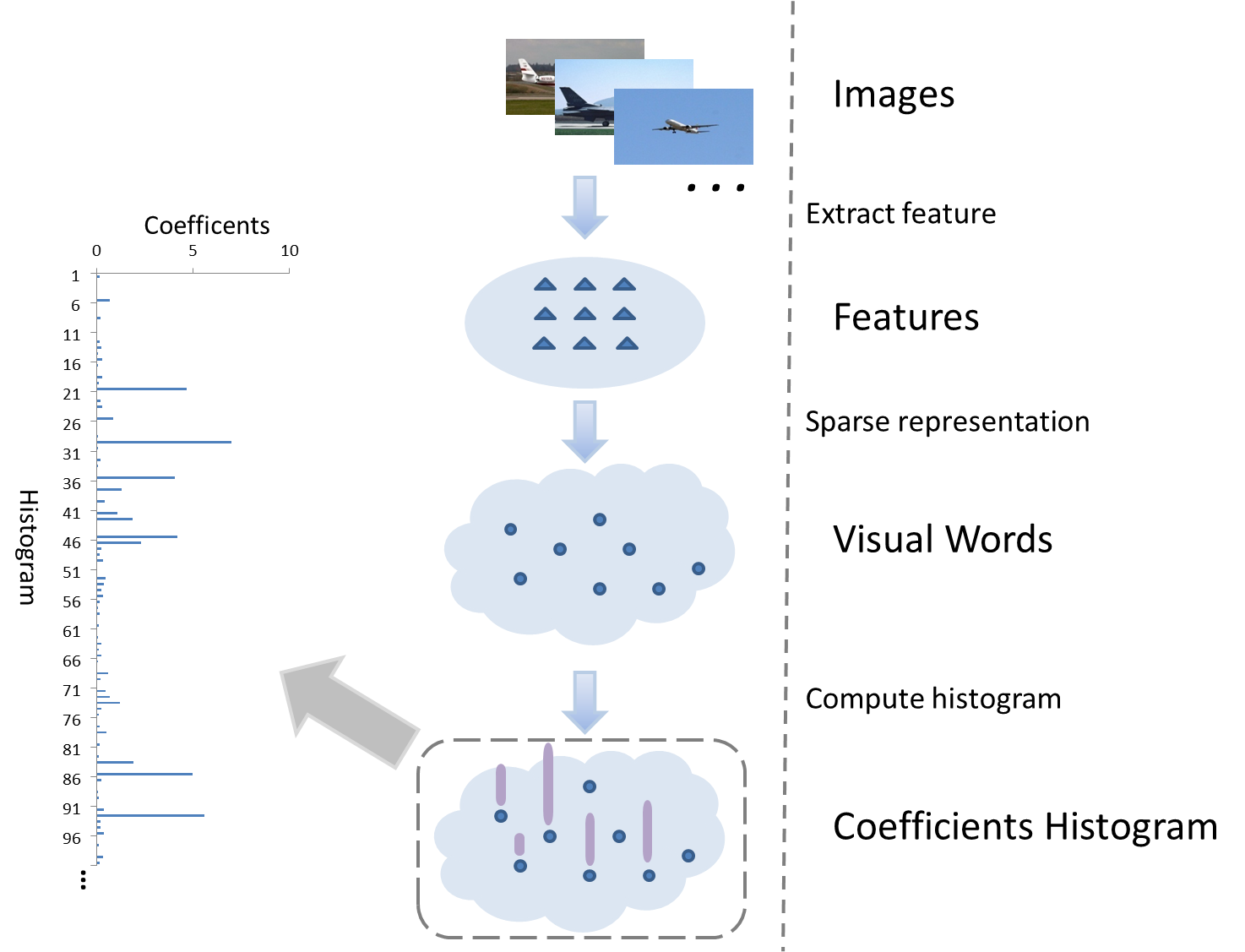


图3.5 BoW模型的实验。图像来自于PASCAL 2007数据库中飞机（Aeroplane）一类。

写清楚公式

如前文所述，稀疏表示具有选择的功效，它能将词典中与样本相类似的词选择出来，并赋予一定的权值。基于这种特征，我们可以考虑使用稀疏表示对BoW模型的码书进行处理，以获取更紧凑且更有区分性的特征表示。

假设从属于类别的图像中抽取出个描述子，记为，这里表示第个描述子，为描述子向量的维度。然后在事先在整个数据库聚类得到的码书上进行稀疏表示：

为表示系数矩阵，它的每一列是第个描述子对应的稀疏表示系数。接下来，我们将所有的描述子对应的表示稀疏相加得到：

然后将系数和通过下面的式子进行量化：

这里是系数和向量的第个分量，并且。取经验值0.1，这能保证大部分的能量被保留下来。指示变量表明了图像的特征向量中那些更加具有区分性的元素，然后将被用作对特征向量的加权。比如，对类别中的一幅图像通过BoW模型抽取出的特征向量为，那么最终的特征将为，这里表示两个向量的元素对应相乘，其结果仍然为一个等长的向量。

写好流程图

下面，我们将对整个算法的框架进行总结：如图3.6所示，上半部分为训练过程，即为前文所述的依靠稀疏表示寻找重要码字的过程，首先对训练样本抽取特征，然后用稀疏表示学习出对码字的加权向量；下半部分则为测试图像的测试过程，类似于BoW模型，首先对图像抽取描述子并量化，接下来使用学习出的权重对量化结果进行加权，然后汲取出最终的特征表示，并送入分类器得出一个置信分数。

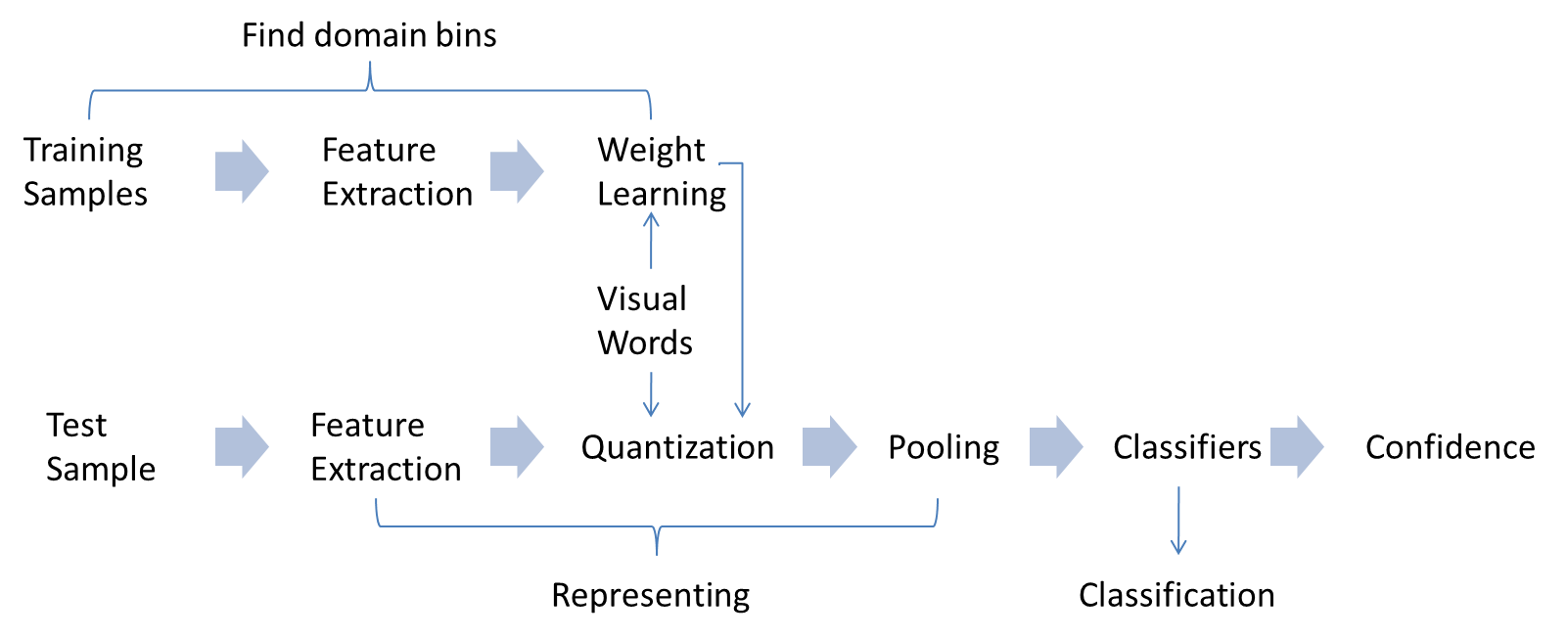


图3.6算法的框架图。

系数表示的意义

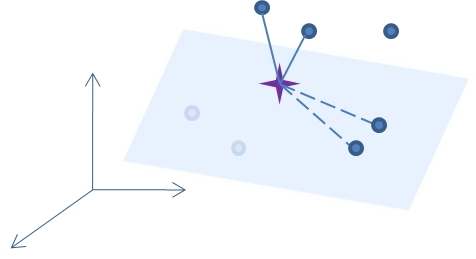


图3.67稀疏表示的意义。

相对于整个码书而言，假设来自于同一类的码字张成了其中的一个字空间。在系数表示中，通过对系数的范数的约束，码书中仅有少量的码字被激活，并且这些激活的码字处在该类码字张成的子空间中。如图3.7所示，一个新的样本（四角星形状所示）被处于同一的空间（平面所示）的样本表示，而不在由其距离的最近的样本来表示。因此，我们认为系数表示可以有效的选择出该类的字空间，这样便使得选择出的码字具有类间的可区分性，得出的特征表示也能更好的代表本类的特性。

3.4 实验分析

先写数据库

前一节我们详细讨论了使用使用系数表示寻找重要码字的方法，我们将在这一节使用实验去验证方法的有效性。

实验是在三个常用数据库上进行的，它们是PASCAL 2007、Caltech-101和Caltech-256。PASCAL 2007数据库总共包含有9,963张图像，其中2501张用做训练集，2510张用做验证集，4952张用做测试集。这些图像分属20个类别，涵盖人物、动物、车辆以及室内等场景。Caltech-101数据库总共含有9,146张图像，分属于101个类别。其中每个类别包含的图像个数从40到800不等，大多数类别含有50张图像。每张图像的尺寸大约为像素。Caltech-256与Caltech-101类似，只是含有更多的图像和类别，约有30,607张图像，分属于256个类别。

再写实验设置

实验的目标设定为图像分类问题，即对于给定的一幅图像，判定其从属于哪一个或者哪几个给定的类别。实验所比较的对象是原始BoW模型以及通过稀疏表示改进后的BoW模型。

实验中，我们采用了SIFT描述子，先在图像中选择出具有局部空间和尺度极值的像素点，然后在一定区域范围内使用具有旋转不变性的描述子进行描述，最后获得了128维的描述子。码书是通过K-Means聚类得到的，其中包含有1000个码字。我们还采用了金字塔匹配（SPM）的方法，这样一幅图像的特征向量含有维。我们使用了支持向量机（SPM）作为分类器，核函数选择了对直方图较为有效的核，松弛变量（Slack Variable）使用3折（3-fold）法在验证集上相互校验而得到。

对PASCAL 2007数据集，我们使用了平均准确率（Average Precision, AP）的准则去衡量系统性能的优劣。这个准则计算了位于准确率/召回率（Precision/Recall）曲线下的面积，较高的AP值则对应着较好的表现性能。对Calteck-101和Caltech-256数据集，我们则采用了分类准确率（Accuracy）作为评判准则。

在PASCAL 2007数据集上实验结果如表3.1所示。我们比较了使用系数表示改进后的BoW模型以及原始的BoW模型。。。。。

在Caltech-101和Caltech-256数据库上的实验结果如表3.2所示。在Caltech-101数据集上，对每一个类别，分别使用了15个和30个样本用作训练，剩余的样本则用作测试。而在Caltech-256数据集上分别使用了30个和40个样本作为训练集，剩余的用作测试。。。。。

表3.1 PASCAL 2007数据集上两种方法的平均准确率（AP）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow |
| BoW模型 | 0.705 | 0.418 | 0.282 | 0.473 | 0.156 | 0.678 | 0.496 | 0.461 | 0.440 | 0.207 |
| 改进的BoW模型 | 0.729 | 0.448 | 0.345 | 0.481 | 0.168 | 0.736 | 0.496 | 0.501 | 0.452 | 0.200 |
| 类别 | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | 平均值 |
| BoW模型 | 0.269 | 0.367 | 0.322 | 0.463 | 0.727 | 0.131 | 0.279 | 0.307 | 0.565 | 0.417 | 0.408 |
| 改进的BoW模型 | 0.399 | 0.420 | 0.364 | 0.519 | 0.736 | 0.142 | 0.306 | 0.313 | 0.635 | 0.457 | 0.442 |

表3.2 Caltech-101 和Caltech-256数据集上两种方法的分类准确率（Accuracy）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | Caltech-101 | | Caltech-256 | |
| 训练样本数 | 15 | 30 | 30 | 40 |
| BoW模型 | 54.81%+0.76% | 64.21%+0.91% | 25.37%+1.22% | 33.16%+1.01% |
| 改进的BoW模型 | 58.13%+0.63% | 66.32%+0.73% | 28.62%+1.30% | 37.46%+1.41% |

3.5 本章小结

本章主要针对BoW模型中所存在的一些问题提出了一种解决办法。在该模型的码书构建过程中，由于没有对噪声加以区分，导致了形成的码书在一定程度包含有噪声。并且由于采用了无监督的聚类的方式，使得形成的码书并不具备类间的区分性。针对这些问题，我们利用稀疏表示寻找码书中那些比较重要且具备类间区分性的码字。由于稀疏表示自身具备用特征选择的功能，可以寻找出那些与待表示样本处于同一字空间的样本，因此可以被用作码字的选择。通过对一类的样本的稀疏表示，我们学习出一个可用作对码字进行加权的指示变量，并将其用作对测试图像的特征向量的加权处理。本章最后，通过在三个广泛应用的数据库上的实验，验证了这个方法可以增强BoW模型的类间区分性，提高了模型在图像分类问题上的性能。

参考文献

第四章分类问题中带有类约束的稀疏表示

引言

先泛泛的讲稀疏表示的一些作用，再引出类约束的好处。然后讲怎么实现的，然后讲通过实验证明了有效性。

在过去的数年中，稀疏表示在高维信号的表示以及压缩方面有着巨大的应用。它的成功主要源于两个方面：一是信号内在的稀疏性，自然界的许多种信号，例如声音或图像，都能在一组固定基上进行稀疏的表示，傅立叶正交基、小波基等都是这些固定基的例子；二是高效解决算法，一些基于凸优化的或者简单的贪婪搜索的算法保证了能在一定界内解决这种稀疏表达的问题。在传统信号处理领域中，稀疏表示的目标是在一定可恢复程度下要求信号的表示更加紧致，更加稀疏。但是在计算机视觉中，我们也许更加关注于图像的内容或者语义等具体的具有物理意义的方面。因此，人们也许会有这样的疑问，稀疏表示能否被用于视觉领域，或者是怎才能应用在视觉领域？生物学给出了一个比较合理的解释，灵长类动物在处理视觉信号时，其视觉系统中的V1区神经细胞仅有少量被激活，这样就能够使用较少的资源编码尽可能多的信息，但更重要的是这种生物机理与稀疏表示是完全一致的。

当然，这个问题也被无数的稀疏表示在计算机视觉中应用所回答，比如，人脸识别，图像超分辨率分析，动作分割，图像降噪和修补，背景建模和图像分类等等，而且这些应用的效果通常都达到了当今最好水平。在众多的应用中，一个重要的思想就是使用这些稀疏的表达系数作为该信号的再表示，然后再将其应用在后续处理中。例如，在人脸识别的应用中【】，一幅测试图像在一个由训练图像组成的基上面进行稀疏表示，由于这组基是带有类标签的，因此可以使用那些稀疏系数在每一个类别对应的基上面尝试重构该图像，最后将其归为具有最小重构误差的类别。直觉上，属于同一个类别的信号之间具有更高的相似性，他们之间能够更好的被彼此表示，因此如果测试信号仅仅在属于同类别的信号上进行表示，那么这种表示将能更好的恢复出原测试信号。基于这种假设，我们在本章提出了带有类约束的稀疏表示，期望能够更好恢复出原有信号，并且得到更好的分类效果。

在对原始信号稀疏表达的基础上，我们对不是从属于本类的表示系数进行l\_2惩罚，然后采用了梯度下降法完成了该问题的求解。我们将在接下来的2，3小节就该问题进行详细的阐述，第4小节对提出的方法进行实验证明及分析，并在最后一个小节给出本章的参考文献。

问题描述

先讲老问题，再讲新的想法。

模型求解

讲怎么样求解的

实验分析

参考文献

第五章 属性过滤文中的低秩

第六章 目标检测中的低秩表示

第七章 总结和展望