一种图像背景视觉语义模板的生成方法*)

王惠锋　孙正兴
(南京大学计算机软件新技术国家重点实验室　南京210093)
(南京大学计算机科学与技术系　多媒体技术研究所　南京210093)

A Visual Semantic Template of Background for Scene Image

WANG Hui-Feng　SUN Zheng-Xing
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)
(Department of Computer Science and Technology, Institute of Multi-Media Technology, Nanjing, 210093)

Abstract Interpreting background is very important in natural scene images. This paper addresses the automatic classification of background region by using visual semantic template. The method to create the templates is introduced first. Then we use these templates to classify background regions, and the results are analyzed. We also use these templates to locate background objects in images, and to determine whether an image contains certain kind of background. The result is promising in object locating. Some approaches to improve the ability of these templates are also discussed.

Keywords Image retrieval, Image segmentation, Background classification, Semantic retrieval

1. 引言

图像低层的物理视觉特征与人的高层认识之间不存在明显的直接联系，这就是视觉信息处理中的“语义鸿沟[1]”，这使得基于图像全局特征的检索结果与人的主观感觉大相径庭。要缓解“语义鸿沟”问题，一个直接的方法是在低层的视觉特征和高层的主观语义之间建立多个中间处理过程[2]，使得两者能够有个渐进的过渡。这种分而治之的策略，需要保证每一步处理结果都要更加有利于主观语义的辨认，同时这些处理流程应该是自然的。根据人观察图像的特点，对于一幅图像，我们主要关注其中的前景对象和主要背景区域。确定图像中包含的对象，是图像理解和语义检索的最关键和最基本的基础之一。目前在图像前景对象的识别方面，研究者已经有了大量的研究，而对于图像的背景对象，却没有得到应有的重视。

一幅图像的内容，不是各个区域的语义的一视同仁的简单叠加。前景和背景作为两个层次的语义，对于不同的应用起着不同的作用。对于自然场景图像来说，图像的背景构成了图像的主要成分，所以尤为重要。在此基础上，可以处理对象的空间关系，场景语义与对象出现间的联系等等，这些都是获取图像的整体语义的自然的过程。

对于前景对象来说，在不同的环境、光照等情况下，其特征的变化非常大。特别在不同的视角，其形状没有一定的规律。直接根据二维的图像来识别前景，难度非常大。背景对象与前景对象的一个很大的不同点是，背景往往是形状无关的，并且在不同的情况下，其颜色和纹理特征具有一定的稳定性。不过，背景区域的位置，可能对于某些背景，比如 sky, water 等，有一定的区分作用。

研究者采取了各种方法来确定户外图像中包含的背景类别。Town[3]使用了多种神经网络，对于多种背景对象进行分类。Campbell[4]首先使用多层感知机，来分类每一个像素属于天空、植物、道路或其他，再用上述结果作为进一步区域分类的上下文，使用神经网络来精确判定一个区域属于更具体的11种对象中的哪一类。上述方法，都要人工对于分类器进行参数调整。本文使用了一种自动的对象分类方法，不需要人工介入。同时，本文的方法对于图像中对象的定位和假设检验问题，也做了一些较深入的探讨。

本文利用了一个包含3400个图像的图像库，所有的图像均从 corel 光盘图像库[5]中选出。图像都是户外的彩色照片，图像的大小是384 * 256或者256 * 384。

2. 图像背景信息的获得

我们使用了一种半自动化的方法来获取图像中包含的多种背景类别。背景类别的获取必须在图像分割的基础上。但是完全自动的图像分割无法取得非常理想的结果，并且不同的分割方法，产生的结果也有很大差别。在不同的分割结果上做背景分类，其结果显然没有可比性。通过人工干预，生成与人的观察一致的图像分割，可以使背景分类有一个一致的基础。问题是如何尽可能让机器自动完成分割任务，而使人工干预尽量少并且有效。

我们首先使用多种自动图像分割方法，并且某一种方法利用不同的参数，来生成多种分割结果。对于每一幅图像，人工选择最好的一种分割。

本文采用的图像分割方法主要有两种：基于子块聚类的分割和 JSEG[6]图像分割方法。

基于子块聚类的方法，首先将图像划分成16 * 16的小块，然后提取每一个小块的特征，通过聚类分析和一系列后处理来生成图像分割。由于使用了子块来聚类，最后生成的分割结果实际上是图像区域的粗分辨率表示。根据聚类使用的子块特征的不同，可以生成多种分割结果。

JSEG 是一种非监督的彩色纹理区域分割方法。整个处理流程分成两个独立步骤：颜色量化和空间分割。同样根据一开始设置的参数的不同，一幅图像可以获得多种分割结果。

*)本文得到国家自然科学基金 (69903006)、教育部高等学校骨干教师资助计划[教技司(2000)65号]和中国博士后科学基金(中博基[1997]11号)资助。王惠锋　硕士研究生，主要研究方向为图像理解、图像检索、视觉信息挖掘。孙正兴　博士后，副教授，主要研究方向为视觉信息处理、多媒体辅助工程。

在图像分割的多种结果中，人工选择一种效果最好的分割。然后在此基础上，人工修正分割，生成与人的观察相一致的结果。

基于子块聚类的分割结果是图像区域的粗分辨率表示，即每一个子块的区域归属。子块表示的粗分辨率区域表示，一个内在的优点是便于人工干预，可以精确指定任一个子块归属的区域。JSEG 分割出来的是基于像素的区域轮廓，我们将其处理成16 * 16的子块的粗分辨率区域，这样便于统一处理。背景分类是基于提取出的子块特征来进行的，图像的子块相对于像素，更便于提取纹理等特征。并且粗分辨率的表示，在处理速度上要优于像素表示，而效果基本接近。

我们设计了一个专门的工具，用来对于子块表示的区域进行操作，可以进行区域的合并，分割，子块的区域指定和删除等。在这个工具的帮助下，我们就可以获得与人的观察相近的区域分割结果。

在经过以上分割处理的图像库中，挑选了 sky，water，grass，tree，snow 五种背景对象进行分类识别。挑选以上种类的背景的原因，一是这些背景在人的视觉上容易定义，二是这些类别在自然场景中出现的频率比较高。

3. 视觉语义模板的原理

同一类自然背景，在视觉特征上往往具有一定的聚合性。比如对于 grass 这一类背景，一般是绿色或是黄色的，并且具有典型 grass 的纹理特征。同时由于光照，反射，聚焦和视角等原因，在不同图像中的同类背景对象的区域，其特征也有较大的变化。所以对于形状特征不敏感的一类背景区域来说，其颜色和纹理特征往往在特征空间上聚合成多个区域。

针对自然背景区域的这种特点，我们使用混合多维高斯分布模型来对它建模。颜色和纹理特征空间上的一个聚合区域，对应于一个多维高斯分布。我们将这种多维高斯分布称为模板。同一类背景，根据其特征变化复杂度的大小，用多个模板的组合来表示，即为混合多维高斯模板。实际上这里要进行的是一次聚类分析，模板的个数就是聚类的数量，模板的参数就是类别的参数。由于这种模板是建立在视觉特征上的，并且用来表示一个语义对象，因此称其为视觉语义模板。

一旦确定了混合多维高斯分布的参数，类别的模板也就确定了。这里的参数主要包括两部分，模板的数目，每一个多维高斯分布模板的密度方程的参数。要使这些参数合理地表示一类背景对象特征的分布，以上的参数都要用大量的样本来训练确定。

在确定了模板数的情况下，要获得混合多维高斯模型的参数，一个适当的方法是使用 Expectation Maximization (EM)[7]。与参数估计的最大似然法相似，它也计算概率密度的最大似然值，但是由于概率密度是混合密度，因此只能通过求混和密度的期望的最大似然值来估计未知参数。

一类背景对象对应的模板数，使用 Minimum Description Length(MDL)[8]来估计。对于模板数目固定的情况，求得其参数，同时求其 MDL。对于不同模板数的情况，MDL 较小者，一般其模板数合理地反映了真实的类别分布，所以取最终的模板数为最小 MDL 对应的模板数。

本文的实验利用了 Cluster 软件包[9]，整个处理过程是：首先选择一个较大的类别数目，估计每个模板的参数，并计算 MDL 值，然后将类别数减1，再次估算上述各值。如此不断循环，直到类别数为1。最后在所有的估计之中，挑选 MDL 最小的那组参数作为最终的参数值。

对于多类背景对象，每一类都按上述方法，确定其对应的混合多维高斯分布的参数。对于一个未知类别的样本，使用 Bayes 法则来将其分类：计算样本属于每个背景类别的后验概率，将其分类到后验概率最大的背景类别中去。计算后验概率时，类内分布是一个类别的多个模板的混合分布，这一点与普通的 Bayes 分类器有所区别。

4. 特征的提取

对于经过处理的背景区域的每一个子块，分别提取它的颜色和纹理特征。颜色特征提取子块的 HSV 颜色空间的一阶和二阶矩[10]；纹理特征提取子块的对比度和粗糙度[11]。其中 HSV 空间的颜色矩对应于颜色在色度、饱和度和亮度上的分布特征，因此需要先将图像的每一个像素的颜色值转化到其对应的 HSV 值。纹理粗糙度和对比度是与视觉感受相关的纹理特征，它们是在灰度图像上计算的，首先都要将彩色图像转换成灰度图像。

接下来进行的分类器的训练实际上是一个聚类分析，它要求有足够的样本数量。而这个数量与特征向量的维数有关系，维数越大，要求的样本越多。在样本数量根据固定的图像大小和子块大小已经确定的情况下，必须保证特征向量的维数不能太高；而对于维度较高的特征向量，不同的特征之间的相互影响，也会导致类别的分散。虽然颜色直方图在检索中一般效果较好，但是由于其一般维数都较高，因此在这里并不适合。而对于颜色矩来说，每一阶只有三个值，并且经过实验，确实具有较好的颜色区分效果，因此比较适合于目前的场合。经过实验，我们还看到，一、二阶颜色矩已经可以较好地描述子块的颜色特征，再取更高阶的矩已经没有意义了。同样，粗糙度和对比度都只具有一个值，并且经过实验检验，对于纹理也确实具有较好的区分度。

一、二阶的 HSV 颜色矩(6个值)和粗糙度(1个值)、对比度(1个值)，就形成了8维的子块特征向量。

5. 三种分类方法及其结果

利用上述获得背景类别的方法，我们共获得了个1023个 sky，441个 water，766个 grass，440个 tree，149个 snow，总共2819个背景区域，作为训练和测试分类器的样本。我们使用了三种不同的方法，来训练和测试视觉语义模板的分类效果。

5.1 区域平均值分类

对于每一个区域，求其所有子块的特征向量的平均值，作为分类器的一个样本。另外，我们将区域在图像上半部的面积，除以区域的面积，作为区域在图像中占据的位置特征。这样将位置特征加入到上述区域平均特征向量上，生成了9维的区域平均特征。

使用视觉语义模板，随机抽取一部分区域进行训练，其余的区域测试。根据获得的后验概率，判断一个样本应该属于哪个类别。将以上数据一半用作训练，一半用作测试，一组典型的结果如表1所示。从表中可以看出，数据主要集中在表1的对角线上，而对角线上表示的是成功分类的数目。不同的背景类别，由于其本身的特点，其错误识别出的对象也有一定的规律。比如 tree 错误识别为 grass，和 grass 错误识别为 tree 的个数就比较多，说明它们都作为植物，相互之间确实具有较高的相似性。另外，由于 sky 区域总数比较多，snow 区域总数比较少，因此 sky 的整体分类精度就比 snow 要高。

表1显示的是第一次就识别成功的精度。我们使用的分类器，一个测试样本对于任何一类背景，都能获得一个后验概率。将样本的多个后验概率排序，最大后验概率对应的类别是第一可能的类别，第二大后验概率是第二可能的类别。这样可

以获得前二次和前三次识别成功的累计精度，如图1所示。从图中可以看出，虽然第一次识别率不是特别高，但是前二次的累计识别率已经在0.95左右。对于一个区域，我们可以将其对应于前两次识别出的两种背景类别，这样对于检索等任务来说，可以尽量避免漏掉相关的图像。同时可以通过相关反馈，来确认图像中区域对应的真实对象类别。

表1 区域平均分类：一组典型结果的混淆矩阵

	tree	sky	Snow	grass	water
tree	184	1	0	26	6
sky	0	480	6	1	14
snow	0	7	56	1	10
grass	14	4	0	347	11
water	8	16	3	26	164

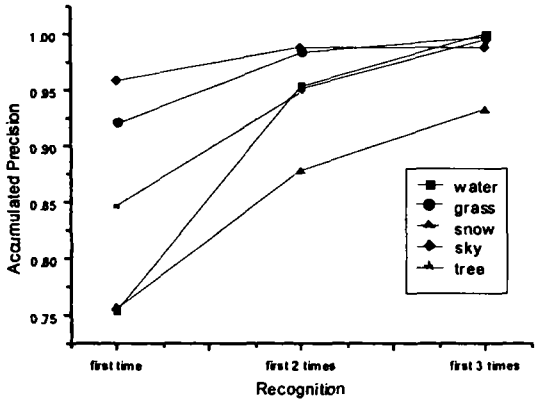

图1 前三次识别累计精度对照表

表2 训练样本数与分类精度对照表

	sky	water	grass	snow	tree	total
1vs3	0.9343	0.7730	0.8513	0.7162	0.8410	0.8602
1vs1	0.9441	0.7788	0.8969	0.7196	0.8353	0.8764
3vs1	0.95	0.7963	0.8497	0.6824	0.7940	0.8600

改变训练集的大小，当训练集和测试集的大小分别是1比3，1比1，和3比1的情况下，5种背景类别的分类精度和所有类别总体的分类精度分别如表2所示。从表中可以看出，不同大小的训练集，其识别精度差别不大。说明每个背景类别本身在内部特征上具有一定的稳定性。另外，也可以看出视觉语义模板，对于训练样本数量的变化具有一定的稳定性。

5.2 子块统计分类

在这种情况下，判断一个背景区域的类别，是通过统计其每一个子块的归属情况来判断的。将同类区域的子块作为样本，训练视觉语义模板。对于一个未知背景区域，利用模板判断其每一个子块的类别。统计所有子块归属的类别，整个区域的类别对应于最多子块归属的类别。随机抽取一定比例的子块来训练模板，然后测试它在所有区域上的效果。子块统计分类的结果和5.1中区域平均分类的结果比较，如表3所示，其中区域平均分类结果利用表2中训练样本数和测试样本数为1比1的情况。从表中可以看出，子块统计分类，在各个单独的背景类别上，以及所有类别总体上的分类精度，都与区域平均分类的结果相近。说明区域子块的平均特征，在一定程度上已经反应了区域整体的特性。但是这里子块统计分类用到的子块数量，平均只占所有子块总数的1/7。所以在区域数目比较少的情况下，子块统计分类比区域平均分类更有优势。

表3 子块统计分类与区域平均分类精度比较

	sky	grass	snow	tree	water	total
Block Statistic	0.9737	0.9247	0.7072	0.8169	0.7558	0.8875
Region Mean	0.9414	0.8969	0.7196	0.8353	0.7788	0.8764

5.3 结合场景上下文的图像背景分类

将区域从图像中单独提出来，判断其所属的类别，对于人来说都不是容易的任务[3]。离开图像中的上下文，来判断一个区域的类别，在某种程度上是不可能完全实现的。只有结合了整个图像的上下文，才能去除一些情况下的歧义性。不同场景类别的图像，其可能包含的背景有很大差别。例如一般的远景的海上图像不太可能包含 grass，tree 和 snow，而 sky 和 water 出现的可能性比较大；而草原和沙漠等图像中一般不会出现 water。如果预先知道图像场景的类别，对于背景对象的分类会有很大好处。对于图像整体的场景信息，既可以从图像本身获得，也可以从与图像关联的外部信息获得。Vailaya[12]利用一系列2类 Bayes 分类器的组合，通过全局的低层视觉信息来获取简单的场景分类信息。他的分类是基于度假的照片：首先区分一个图像是室内的还是室外的，再对于室外的图像区分它是城市中的还是野外风景；然后对于野外风景再区分是日落还是森林或者山地。每一个2类分类器都评价各种全局视觉特征对于特定的2类问题的区分能力，选取出最好的一种或几种特征的组合。根据他的实验，空间颜色和亮度分布对于区分 indoor 和 outdoor 问题比较有效，边的分布对于区分 city 和 landscape 比较有效，全局颜色分布和饱和值对于区分 sunset，forest 和 mountain 比较有效。这些2类 Bayes 分类器的层次式组合，对于简单的环境分类，具有较高的分辨率。下面第6部分还将讨论从外部获取这些信息的可能性。

本文所利用的 corel 图像库，每一个目录对应于一个图像分类，有100幅图像。我们首先计算 sky，water，tree，grass，snow 在每一个图像分类中出现的频率；然后在5.1节中区域平均分类的基础上，结合统计出来的频率，来确定一个背景对象属于某类背景的可能性。比如计算一个图像中某个背景对象属于 sky 的可能性，直接把5.1节中获得的该背景对象属于 sky 的后验概率，乘以该图像所属分类中 sky 出现的频率，作为衡量的标准。一组典型的结果，及其相应未使用图像场景上下文的情况，如表4所示。从表中可以看出，各个背景类别的识别正确率，分别有不同程度的提高，总体的识别率提高了4%。对于 snow 这种主要出现在少数场景中的背景类别，其识别率提高得最多，超过10%。说明在背景分类时结合一定的场景上下文，对于背景分类的精度确实是有作用的。

表4 结合场景上下文与未使用结果对照表

	sky	grass	snow	tree	water	total
With Context	0.9681	0.9295	0.7703	0.8399	0.8007	0.9007
Without Context	0.9247	0.8910	0.6655	0.8306	0.7615	0.8614

6. 背景对象定位

在图像中存在其他各种区域的情况下，直接利用有限的类别模板在任意图像中识别某个区域的类别，还有一定的难度。目前的一个比较实际的应用是，在图像中已知某种背景对象存在的情况下，将其对应的具体区域找出来。在某个区域上，标上对应的类别标签，是普通的区域标注。而已知图像中出现的类别，将该类别对应到具体的区域，与普通的标注过程相反，称为对象定位。

图像中包含的对象信息，可以从各种外部信息源来获取。由于多媒体描述方式的普及，目前图像出现在各种场合。报刊杂志、论文报告等电子文献，特别是在 Internet 上的各种 Web 网页中，都出现大量的图像。Internet 上的图像，和出现在同一 Web 页面中的其他的信息之间具有很大的关联。关系较大的内容主要包括 URL 中的文件名、IMG 的 alt 域、图像前后的文本等。从这些地方可以获得图像中包含的对象的可能信息。另外传统的图像检索，有时也对于整幅图像先做了标注，这里面就有图像包含的对象信息。对于只做了场景的标注，而没有指明图像包含的具体对象的情况，可以统计一类场景中经常出现的对象类别，作为图像包含对象的假设情况。

对于明确知道了图像中已经包含某一类背景的情况，利用上述的背景模板，从图像的所有区域中找出与该类模板最相似的区域，作为这种背景对应的区域，这个过程称为定位。对于图像中可能存在某种对象的情况，也用该类对象的模板去与所有区域匹配，找出最相似的区域。如果相似度大于一定的阈值，就认为该图像确实包含该类区域，并且也确定了区域的位置。如果相似度不够大，则认为图像中不包含该种背景（拒绝），将这个过程称作假设检验。

表5 图像中的对象定位正确率

	sky	grass	Snow	tree	water	total
All Images	600	417	76	265	263	1621
Success	582	358	73	223	216	1452
Fail	18	59	3	42	47	169
Precision	0.97	0.8585	0.9615	0.8415	0.8217	0.8957
Error Rate	0.03	0.1415	0.0385	0.1585	0.1783	0.1043

利用上述试验数据，做了如下实验。对于某个图像，已知其肯定包含了上述5种背景的一种，利用背景模板定位这个背景具体对应的区域。利用一半的区域样本训练模板，另外一半区域对应的图像做测试，对象定位的一组典型的结果如表5所示。从表中可以看出，总体的定位正确率接近90%。对象定位在图像检索中的一个典型应用是：用户在检索结果中选择了若干幅图像，并且对于这些图像统一标注了一种背景类别，这样就可以使用对象定位，将标注细化到区域的粒度级别上。这可以看作通过用户反馈来生成和细化图像语义的一种方法。

在上述对象定位的基础上，加入了拒绝机制。采用与表5实验相同的模板和同样的测试图像，通过实验设定了一个适当的阈值，假设检验的结果如表6所示。从表中可以看出，在12.21%的拒绝率下，对象定位错误率降低了3.15%。

表6 对象定位中引入拒绝后的结果

	sky	Grass	snow	Tree	water	total
All Images	600	417	76	265	263	1621
Success	538	339	65	173	190	1305
Fail	9	51	2	17	39	118
Refuse	53	27	9	75	34	198
Refuse Rate	0.0883	0.0647	0.1184	0.2830	0.1293	0.1221
Error Rate	0.015	0.1223	0,0263	0.0642	0.1483	0.0728

表7 不包含相应背景的纯粹拒绝率

	sky	grass	snow	tree	water	total
All Images	547	730	1071	882	884	4114
Refuse	467	474	808	707	385	2841
Fail	80	256	263	175	499	1273
Refuse Rate	0.8537	0.6493	0.7544	0.8016	0.4355	0.6906

下面我们再测试肯定不包含某种背景对象的情况下的假设检验的效果。用表5和表6实验相同的测试图像集，如果一幅图像肯定不包含 sky，则将其作为 sky 的假设检验测试集，其他情况依次类推。实验结果如表7所示。从表中可知，在当前的阈值下，能对肯定不包含相应背景的图像，总体达到接近70%的拒绝率。如果将阈值加大，则拒绝率会进一步增高，对于不包含相应背景的图像，结果越好。但是对于包含相应背景的图像，虽然其错误率还会进一步减少，但是由于拒绝率的增高，其正确定位的图像总数也会减少。所以阈值的设定必须在权衡正确定位和正确拒绝的基础上设定。不同的类别，其混合多维高斯密度分布是不同的，所以不同的类别应该设定不同的阈值。上述实验中对于所有的5类背景采用了统一的阈值，所以对于不同的背景类别，其结果差别比较大。通过实验选择每一类的阈值，将会对整个结果有一定的提高。

结束语 要使图像检索和图像理解进一步接近人的观察机制，找出图像中的对象是基础，确定图像包含的背景也是重要的一个方面。对于同一类的背景，其视觉特征具有一定的不变性。本文利用这种不变性，构造了背景的视觉语义模板，在一定的情况下可以利用这种模板来分类图像中的背景区域。另外，本文利用背景视觉语义模板进行了对象定位，在已知图像包含某类背景的情况下将背景对应到具体的区域；或者假设图像包含某类背景，检验这种假设是否成立。从实验结果来看，对象定位和假设检验可以取得很好的效果。目前的方法对于所有的背景类别，使用了固定的几种特征来构造视觉语义模板。如果针对不同的背景，利用特征选择方法，选出区分能力最强的几种特征，以及相应的特征匹配方法来构造模板，应该能取得更好的效果，这也是进一步研究的方向。

参考文献

1 Gudivada V N, Raghavan V V. Content-based Image Retrieval System. IEEE Computer, 1995,28(9): 18～22

2 王惠锋，孙正兴. 基于内容的图象检索中的语义处理方法. 中国图像图形学报，2001,6(10):945～952

3 Town C P, Sinclair D. Content Based Image Retrieval using Semantic Visual Categories. http://www.uk.research.att.com/permm/,2002

4 CampBell N W, Mackeown W P J, Thomas B T, et al. Interpreting Image Databases by Region Classification. Pattern Recognition, 1997,30(4):555～563

5 Wang J Z, Li J, Wiederhold G. Corellm database used in SIMPLIcity. http://wang.ist.psu.edu/docs/related/, 2000

6 Deng Y, Manjunath B S, Shin H. Color image segmentation. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 1999.446～451

7 Bilmes J. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation Gaussian Mixture and Hidden Markov Models. http://citeseer.nj.nec.com/bilmes98gentle.html, 1998

8 Rissanen J. Modeling By Shortest Data Description. Automatica, 1978,14:465～471

9 Bouman C A. An Unsupervised Algorithm for Modeling Gaussian Mixtures. http://dynamo.ecn.purdue.edu/～bouman/software/cluster/, 2000

10 Stricker M, Orengo M. Similarity of Color Images. Proc. of SPIE-Storage and Retrieval for Image and Video Databases III, 1995,2420:381～392

11 Hideyuki T, Shunji M, Takashi Y. Textural Features Corresponding to Visual Perception. IEEE Trans on Systems, Man, and Cybernetics, 1978, SMC-8(6):460～473

12 Vailaya A, Figueiredo M, Jain A K, Zhang H J. Image Classification for Content-Based Indexing. IEEE Trans. on Image Processing, 2001,10(1):117～130