一种新的个性化的图象分类方法

李金龙 王上飞 陈恩红 许月华 王煦法
(中国科学技术大学计算机系，合肥 230026)

摘 要 图象分类系统的建立是信息检索以及模式识别中一个重要部分，其中，特征选择问题，即确定描述图象的特征参数是需要解决的关键问题. 基于内容的图象检索技术的研究，近来得到了广泛的关注，由图象特征向量维数过高而引起的图象检索困难是基于内容的图象检索技术研究所面临的一个挑战，因此需要寻求一个有效降维技术. 为解决此问题，设计了一个新的图象分类标准模型，通过寻找不同的特征组合来作为分类标准，进而提出了一种算法，用于实现此模型. 实验结果显示，该模型能实现图象特征向量降维，并且算法能够极大地降低计算所花费的时间. 同时，多种不同分类标准的引入，使得本方法能与信息检索技术进行有效的结合，为个性化信息检索提供一种实现思路.

关键词 多标准图象分类 降维 个性化
中图法分类号：TP391.41 TP183 **文献标识码**：A **文章编号**：1006-8961(2002)11-1156-05

A New Personalized Image Classification Method

LI Jin-long, WANG Shang-fei, CHEN En-hong, XU Yue-hua, WANG Xu-fa
(*Department of Science and Technology of USTC, Hefei* 230026)

Abstract Image classification system is an important part of any information retrieval system and pattern recognition system, and its key issue is to select some appropriate feature bindings of an image. Recent years content-based image retrieval has been a very active research area. The dimension of the image feature vectors is normally very high and it's hard to index images. One of the main challenges in content-based image retrieval is to develop techniques of performing dimension reduction. In this paper, a new model of searching multiple classification criterions has been proposed in which different feature bindings were formed to find new classification criterions, and a new algorithm was designed for this model. The experimental results shown that the proposed model can perform dimension reduction. The algorithm for the model is capable to reduce computational time which was also illustrated with results. The multiple criterions in combination with the information retrieval techniques can implement personalized information retrieval, and some results were given in last section.

Keywords Multiple criterions image classification, Dimension reduction, Personalization

0 引 言

随着 Internet 的发展，越来越多的文档(包括文本、图象、音频和视频等)出现在 Internet 上，形成了一个巨大的分布式文档库. 这样，用户为了在 Web 上找到满意的信息，往往要耗费大量的时间和精力. 这使得网上信息检索技术在近年受到了人们的广泛关注. 鉴于网上信息中大量的是多媒体信息，因此多媒体信息检索技术的研究就成为当前研究热点[1,2]. 一般多媒体信息检索包括对图象、音频、视频信息的检索等，其主要特点之一是由多媒体信息的直观性、多义性而导致的个性化检索[2]. 而对多媒体信息的分类则是实现网上信息检索的关键问题之一.

分类是模式识别中的一个重要问题，而其中的特征选择又是任何一个分类系统均需要解决的一个重

基金项目：国家 973 计划项目(G1998030500)
收稿日期：2001-07-23；**改回日期**：2002-01-10

要环节,其主要目标是在获得最优、最显著有用特征的同时,丢弃无关或次要的信息,以降低分类系统的复杂性[3]. 而特征选择最难的问题是评估每一特征在分类系统中的作用,即如何判断特征在决策过程中是否相关或是否是次要的. 迄今为止,人们在这一方面作了不少工作[3~5],其中大部分是先定义一种对特征的评估标准,然后依标准将特征进行排序;或者由用户通过直觉寻找最优特征[5]. 本文认为,最优特征不但很难确定,而且其分类结果对用户而言,却不一定是最好的,因为不同的用户对同一个分类结果有不同的理解. 因此本文为满足不同用户的要求,提出利用多种特征组合来对图象进行分类,并根据用户的需求来选择分类方式. 目前,国际上已开始讨论如何将多个分类器进行组合,以获得良好的分类性能[6~8]这一问题,但这些工作的立足点仍是寻找最优特征或特征组合,并没有涉及如何获得适用于多用户的分类标准和方法,国内也尚未见到有关这方面工作的报导.

从个性化和适用多用户分类的角度出发,本文设计了一个新的图象分类标准模型,这样通过寻找不同的分类标准,即可实现对图象的自动分类和在检索系统中实现降维索引. 为建立分类标准模型,本文提出一种多标准自动分类算法 MCCA(Multiple Criterions Classification Algorithm),用来寻找不同的分类标准,并将之用于基于内容的图象分类. 实践证明,MCCA 与信息检索技术的结合可有效地实现个性化检索.

1　分类标准

假定多媒体特征抽取完成后,可得到一个 N 维的特征空间,而对任一 $n(\leq N)$ 维多媒体信息,就有 n 维特征向量 $(a_1,a_2,\cdots,a_n)$. 可见,对多媒体信息(如对 M 幅图象)进行分类,实际上是对 m 个特征向量 $(a_1,a_2,\cdots,a_m)$ 进行分类,即依照特征向量中每一分量 a_i 的数值大小或多少,将这些特征向量归类. 由于多媒体信息的多义性,不同的人对同一信息的理解和看法不同,例如,某些人对分量 a_i 感觉更敏锐或更加关注分量 a_i,并不在意或完全忽略其他分量,而另外一些人则可能与此相反,所以不能将所有图象简单地按一个统一标准来分类,而需要将图象按照多个不同的分类标准进行分类.

为简化问题,假定用户对特征分量 a_i 的关注程度分为两类,分别用 1(关注)和 0(不关注)来表示,并且特征分量 a_i 在 n_i 个离散值中取值. 下面给出特征组包含的定义:

定义 1　给定 N 维特征空间 Ω 的两个投影 Ω_{proj1} 和 Ω_{proj2},如果对任意分量 $a_i\in\Omega_{proj1}$,并有 $a_i\in\Omega_{proj2}$ 成立,那么称 $a_i\in\Omega_{proj2}$ 包含 $a_i\in\Omega_{proj1}$.

分类标准模型可描述为:给定一个 N 维特征空间 Ω,其中有 n_Ω 个点,构成特征点集合 A,其求出的所有 K 维投影子空间 Ω_{proj},即用户关注的特征空间,但需满足如下条件:

(1) $1\leq K<N$

(2) $n_{\Omega_{proj}}/n_\Omega<c$,$c$ 是一常数,n_Ω 为特征空间 Ω 中要分类的特征点数量,$n_{\Omega_{proj}}$ 是特征空间 Ω 中的特征点经过投影后所剩的特征点数量.

投影子空间 Ω_{proj} 中的一个点是原特征空间 Ω 中一个或多个点的投影,条件(2)指出,当 c 取小的值(<0.1)时,由于 Ω_{proj} 对 Ω 中的点具有很好的聚合能力,因此称 c 为聚合常数. 这种聚合能力表明,Ω_{proj} 的基所代表的特征组合是对图象内容的进一步抽象,具有分类意义,是可能的分类标准. 因为模型中的问题具有 $O(2^n)$ 复杂度,并且条件(2)中的聚合常数是个待定参数,难以处理,所以本文提出用 MCCA 算法来建立分类标准,并对图象进行自动分类.

2　MCCA 算法框图及步骤

MCAA 算法分类过程如图 1 所示:

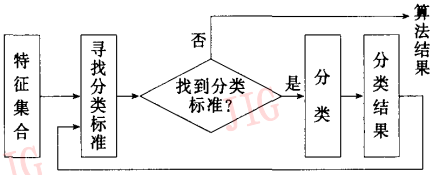

图1　算法框图

算法由寻找分类标准及分类两个核心过程组成,各个过程的详细描述如下:

(1) 寻找分类标准的过程. 如果 Ω_{proj2} 包含 Ω_{proj1},那么 Ω_{proj2} 就一定由 Ω_{proj1} 生成. 本过程其核心思想就是由 $K-1$ 维 Ω_{proj} 来生成 K 维 Ω_{proj},因为这样可以裁剪掉大量不合要求的搜索分支. 寻找分类标准过程是,首先计算特征或特征组在特征向量中出现的频度 $S=m/M$,其中,m 为特征或特征组在全体向量中出现的次数,M 为向量总数. 由于出现频度高的特征或特征组,其聚合能力 $c=n_{\Omega_{proj}}/n_\Omega$ 的值较小,因此,MCCA 算法中,利用频度阈值 s 这个实验常数来替

代聚合常数 c,两者之间的关系可用 $s=\alpha/c$ 来描述,由于实际实验过程中,α 和 c 都是未知的,且 α 是和向量总数相关的一个常量,因此,寻找分类标准过程就转化成为寻找频度高的特征组合的过程.

(2) 分类过程. 在发现了一条分类标准 Ω_{proj} 之后,再用此分类标准计算特征向量之间的距离,并对全体特征向量进行聚类分类. 分类标准 Ω_{proj} 可转换成与某一频度阈值有关的向量 $\boldsymbol{W}_i=(w_1,w_2,\cdots,w_n)$,其中 $w_i\in\{0,1\}$,当且仅当特征分量 $a_i\in\Omega_{proj}$ 时,$w_i=1$,并将 $\boldsymbol{W}_i$ 作为分类过程中距离计算公式的权向量. 对任意两张图片,其特征向量设为 $\boldsymbol{t}_i=(a_1,a_2,\cdots,a_n)$,$\boldsymbol{t}_j=(b_1,b_2,\cdots,b_n)$,用作分类计算时,度量两个特征向量 $\boldsymbol{t}_i,\boldsymbol{t}_j$ 距离 d 的计算公式为

$$d=\left(\sum_{i=1}^{n}w_i(a_i-b_i)^2\right)^{1/2} \tag{1}$$

具体分类标准寻找算法如下:

输入:频度阈值 s,分类数 v,特征向量集 $\boldsymbol{X}$;

输出:特征向量分类结果,分类标准集 $\boldsymbol{L}$;

(1) 初始化,并输入频度阈值 s 以及分类数 v;

(2) 求出频度大于频度阈值 s 的所有特征 $\boldsymbol{t}$,得到集合 $\boldsymbol{L}_1$,并使得 $\boldsymbol{D}=\boldsymbol{L}_1$;再构造集合 $\boldsymbol{L}_k$,$\boldsymbol{L}_k$ 表示集合中的每一元素是 k 个特征的组合,即集合的每个元素是长度为 k 的特征组

(3) for ($k=2;\boldsymbol{L}_{k-1}\neq\varnothing;k++$) {

$\boldsymbol{C}_k=\{c\in\boldsymbol{C}_k\mid c=l\bigcup\{t\},l\in\boldsymbol{L}_{k-1},t\in\boldsymbol{D}\}$ 表示所有可能的长度为 k 的特征组集

for ($\boldsymbol{C}_k$ 中所有的特征组 $\boldsymbol{t}$) {

if ($\boldsymbol{t}$ 的频度>频度阈值 s) then {

$\boldsymbol{L}_k=\boldsymbol{L}_k\bigcup\boldsymbol{t}$;

用 $\boldsymbol{t}$ 作为图象检索系统的索引

或者

利用公式(1)计算距离,对全体特征向量进行分类}//end of if $\boldsymbol{t}$ 的频度>频度阈值 s

}//end of for $\boldsymbol{C}_k$ 中所有的特征组 $\boldsymbol{t}$

$\boldsymbol{D}$ 为 $\boldsymbol{L}_k$ 中所有特征集合,其中每个元素都是单个特征;

}//end for

(4) $\bigcup_k\boldsymbol{L}_k$ 得到分类标准集合,集合内每一元素代表一个分类标准.

整个算法得到的分类标准保存在 $\bigcup_k\boldsymbol{L}_k$ 中,若有 P 条分类标准,则分类过程将执行 P 次,且所得到的每一种分类结果都以一定的特征作为其分类依据. 在将该算法用于图象检索系统时,因为寻找分类标准发现过程以及分类过程,只需要在图象数据库系统建立和更新时,重新执行,所以并不增加用户在使用图象检索系统时的检索时间,相反,会由于检索系统对图象的进行准确分类,从而不仅可减少用户在检索过程中花费的时间,并且是基于内容的检索,其准确性也得到提高.

3 实验设计及实验结果

基于上一节的思想和算法,本文对图象分类进行了实验. 实验是采用一个拥有32 623张彩色图象的实验数据库中的所有图象. 所抽取的图象特征共有 968 维,其中包括图象的色彩、形状、空间位置和纹理 4 大类. 由于不同类别特征之间具有一定的差异,故本文对每一类图象特征分别用 MCCA 算法进行计算,其中频度阈值 s 的值分别为 0.1,0.01,0.000 1,0.000 01.

图 2 给出了频度 s 和所寻找到的分类标准数目关系图.

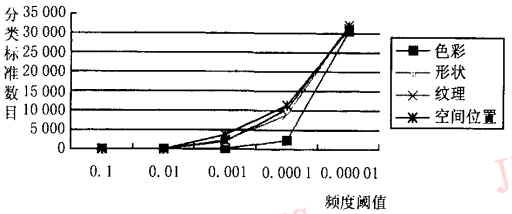

图 2 频度阈值-分类标准数目关系

从图 2 可见,频度阈值 s 越小,所寻找到的分类标准数目越多,但是随着频度阈值 s 的减小,分类标准数目趋向图片总数,这样所发现的分类标准就失去分类意义. 频度阈值 s 可因特征类别不同而不同,当其取值在[0.01,0.000 1]之间,则将获得较合适的分类标准数目.

频度阈值 s 不仅控制着分类标准的数目,并对算法的计算时间影响很大,图 3 描述了频度阈值 s 和计算时间之间的关系. 实验采用的系统配置为 PⅢ766处理器,128M 内存,LINUX 操作系统,算法实现用 C 语言.

图 3 中所示处理时间是应用 MCCA 算法分别处理 4 类特征所花费时间之和. 图中的 not MCCA

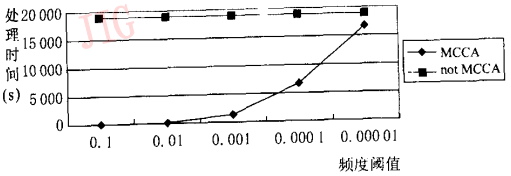

图 3　频度阈值和处理时间的关系

指的是不用 MCCA 算法,而是直接根据本文提出的分类标准模型进行全部子空间测试的处理时间. 由图 3 可以看出,不用 MCCA 算法,处理时间基本保持不变,因为在各种频度阈值下,需要对 n 维空间的全部 2^n 个子空间都进行检测. 由于本文提出的 MCCA 算法的计算时间随频度阈值变化,当 s 的值较大时,则能裁剪掉大量的搜索分支,因此可节省处理时间. 实验表明,当频度阈值 $s>0.000\,1$ 时,MCCA 算法花费时间较少,具有良好的性能.

另外,还注意到聚合常数的取值小于 0.1 时,聚合能力较好. 由图 2 可知,聚合常数大小会因特征类别的不同而不同,故难以处理. 虽然公式 $s=\alpha/c$ 描述了频度阈值 s 和聚合常数 c 的关系,但是由于 α 本身也是个未知数,因此,这个公式也不能准确给出频度阈值 s 或者聚合常数 c 的取值范围,在分类或检索时,合适的频度阈值 s 的选择很重要. 作为实验常数,频度阈值 s 的取值应根据所花费的处理时间和合适的分类标准数目来权衡选取. 本文取定 $s=0.001$,通过对实验结果的分析,利用 $n_{\Omega_{proj}}/n_\Omega<c$ 计算出聚合能力(常数)c 的值约为 $0.07(<0.1)$,实验证明,将所获得的分类标准应用于基于内容的图象检索系统,效果较好.

在图象检索过程中,首先利用用户和系统的交互来统计和学习用户的行为和倾向,然后将用户倾向的分类标准与分类标准集合中的标准进行匹配,并直接利用分类标准作为图象检索索引,这样不仅大大提高了检索速度,也降低了检索的维数. 在检索实验中,用于索引的特征维数平均为 103.3,仅为降维前索引维数的 10.7%. 图 4 给出了一次检索结果. 实验由 3 个不同用户检索具有“蓝天青山”特征图象的结果,其中,图 4(a)表示的第 1 用户的侧重点在优先考虑“天”和“山”的相互位置,而色彩和其他特征次之,用户和系统交互 5 次,系统用匹配出的分类标准集合中的第 492 个标准来对图象库进行检索所得结果;图 4(b)表示第 2 个用户优先考虑“蓝天和“青”地的色彩,而位置和形状等特征次之,用户和系统交互 4 次,系统用匹配出的分类标准集合中的第7 396个标准 GO 对图象库进行检索所得结果,这与第 1 个用户所采用的分类标准集合中的第 492 个标准不同,虽然两用户检索同样的内容. 此实验表明,系统根据用户的不同而调整采用的分类标准,以适应不同用户的特点和需求;图 4(c)表示第 3 个用户检索时,没有利用本文建立的多分类标准,而是考虑所有特征,用户和系统交互 8 次所得的结果. 从人机交互的次数上看,本文的方法平均交互次数为 5.3 次,而不用本文方法,平均需要 9.2 次,可见本文的方法提高了检索速度,同时检索结果的准确度也得到了较大提高. 比较图 4(c)和图 4(a)、图 4(b)可见,因其对所有的特征都进行统一评估,所以有可能在计算中,使重要的特征不能突出其作用,而检索出一些和目标相差很大的图象,如图 4(c)中第 3 张图象. 图 4(a)和图 4(b)的检索结果,清晰体现了两个用户对图象中物体的位置关系和色彩的倾向程度的差异. 重复多次的检索实验结果表明,本文所提出方法的快速和准确.

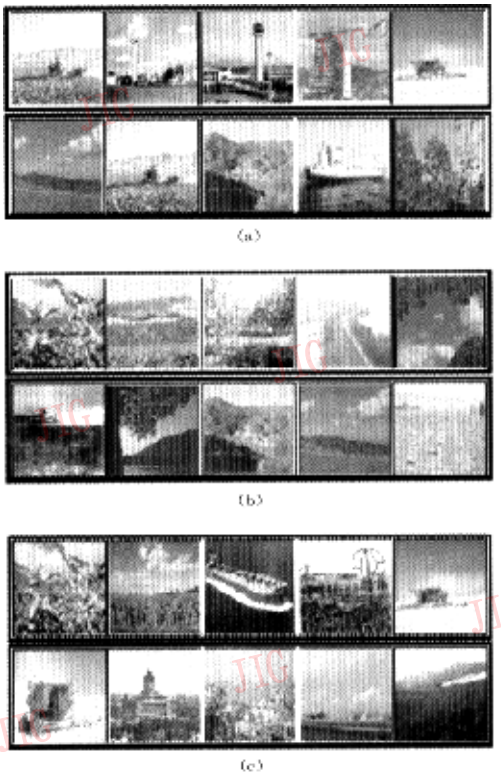

图 4　3 个不同用户的检索结果

4 结束语

为了满足不同用户的分类要求,以实现个性化图象分类和个性化图象检索,本文从个性化分类标准出发,设计了一种新的图象分类标准模型,并提出了一个 MCCA 算法,用于实现此模型,最后将所获得的分类标准用于图象自动分类和基于内容的图象检索.由于采用了多标准分类,因此在基于内容的图象检索系统中,所获得的检索结果不仅对不同的用户具有很强的自适应能力,而且对实现个性化图象检索系统具有重要意义.由于本文提出的方法是基于对信息特征的处理,因此不但适用用于图象自动分类和检索,也适用于其他多媒体信息的分类和检索.对所获得的分类标准继续优化和补充将是本文的进一步的工作.

参 考 文 献

1 Vailaya A, Figueiredo M, Jain A K *et al*. Image classification for content-based indexing[J]. IEEE Transactions on Image Processing, 2001,10(1):117~130.

2 Rui Y, Huang T S, Chang S F. Image retrieval: Past present and future[J]. Journal of Visual Communication and Image Representation, 1999,10:1~23.

3 Vailaya A, Jain A. Reject option for VQ-based bayesian classification[EB/OL]. http://citeseer.nj.nec.com/321657.html,2000-9.

4 Belue L M, Bauer K W. Determining input features for multilayer perceptions[J]. Neurocomputing, 1995, 7(2): 111~121.

5 Basabi Chakraborty, Yasuji Sawada. Feature selection by artificial neural network[A]. In: Methodologies for the Conception, Design and Application of Soft Computing Proceedings of IIZUKA'98[C], Iizuka, Japan, 1998:247~250.

6 Kittler J, Hatef M, Duin R P W *et al*. On combining classifiers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998,20(3):226~239.

7 Jaimes A, Chang S F. Automatic selection of visual features and classifiers [EB/OL]. ftp://ftp.ee.columbia.edu/CTR-Research/advent/public/papers/00/ajaimes-spie00_storage.html, 2000-1.

8 Jaimes A, Chang S F. Integrating multiple classifiers in visual object detectors learned from user input[EB/OL]. ftp://ftp.ee.columbia.edu/CTR-Research/advent/public/papers/00/ajaimes-accv2000_comb.html,2000-1.

李金龙 1975 年生,1998 年开始在中国科学技术大学计算机系攻读硕士和博士学位.现主要研究方向为计算智能、视觉信息处理和复杂系统等.发表论文数篇.

王上飞 1974 年生,博士研究生,1999 年获得中国科学技术大学电子科学与技术系硕士学位.主要研究领域为智能信息处理、图象检索、模式识别.发表论文数篇.

陈恩红 1968 年生,博士,副教授.现主要从事数据挖掘与知识发现、文本及图象信息检索、约束满足问题等领域的研究.发表论文 30 余篇.

许月华 1982 年生,现为中国科学技术大学计算机系硕士研究生.主要学习和研究方向是机器学习、网络信息检索.

王煦法 1948 年生,教授,博士生导师,1970 年毕业于中国科学技术大学无线电系,现为中国科大信息科学与技术学院副院长.主要研究领域为计算智能、智能信息处理、网络安全等.发表论文 60 多篇,出版学术专著 4 部.