基于特征元素和关联规则的图象分类方法

李 勍,章毓晋
(清华大学电子工程系,北京 100084)

摘 要: 图象分类是搜索引擎中的重要模块.本文提出了一种基于特征元素的图象分类方法.特征元素与特征向量相比能够根据人的主观感知来提取图象的视觉特征.与传统的基于特征向量的图象分类方法不同,本文提出的图象分类方法不计算特征空间中特征向量之间的距离,而是通过关联规则挖掘发现图象的特征元素与图象所属类别之间的联系.本文实现了该分类算法并将其与一种基于特征向量的图象分类方法 NFL 相比较.实验的结果证实了所提方法的优越性.

关键词: 图象分类; 特征元素; 关联规则挖掘; 特征向量
中图分类号: TN911.8; TP391 **文献标识码:** A **文章编号:** 0372-2112 (2002) 09-1262-04

Image Classification Based on Feature Element and Association-Rule

LI Qing, ZHANG Yu-jin
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: With the growth of Internet and storage capability in recent years, image has become a widespread information format in World Wide Web. However, it has become increasingly difficult to search for images of interest, So effective image search engine for the WWW needs to be developed. Image classification is an important module in image search engine. In this paper a novel approach for image classification based on feature element is proposed. Compared with feature vectors, feature elements can capture visual meanings of the image according to subjective perception of human beings. By using feature element, our approach for image classification is totally different with traditional image classification method. It does not calculate the distance between two vectors in the feature space, while trying to find associations between feature element and class attribute of the image. After the implementation, we have compared it with NFL a traditional image classification algorithm based on feature vector. Some improved results are presented.

Key words: image classification; feature element; association rule mining; feature vector

1 引言

近年来随着网络应用的普及和存储能力的增长,图象已经成为 WWW 上极为普遍的信息载体.越来越多的人希望能够在 Internet 上检索需要的图象.一些 WWW 上著名的文字搜索引擎如 Alta Vista 等已经开始提供对图象的检索服务,但大部分图象搜索引擎仍是基于关键字的检索,前提是描述图象的文本已经存在.虽然用文字对图象归类对于 WWW 上的图象来说很常见,但是由于基于文本的搜索引擎在图象信息的索引和检索方面的能力非常有限,对于更加有效的图象搜索引擎的研究显得日益重要.

基于内容的图象检索(CBIR)研究近年来一直十分活跃,并出现了一些 WWW 上运用该技术的图象搜索引擎[1,2].CBIR 的主要思想是从图象中提取出诸如颜色、纹理、形状等特征并将这些特征用向量表示,图象之间的相似度就定义为特征向量之间的距离.与基于文本的图象检索相比,基于内容的图象检索能够自动提取特征向量,并且在一定程度上反映了图象的视觉内容.现在 CBIR 研究的主流趋向于基于语义的检索,希望通过各种特征的结合来获取人对图象的语义认知.然而在特征向量的基础上提取图象的语义十分困难,因为特征向量并不能真正地反映人对图象的理解.例如当一个人看一幅彩色图象时,很难认识到该图象的颜色直方图,而更可能关心的是这幅图象包含了几个特定的颜色.视觉心理学的大量实验结果表明,人们对图象内容的认知是离散的而非连续的.特征元素的理论就是在这个基础上提出的[3].与向量空间中连续的特征向量不同,特征元素采用了离散的形式.实质上,特征元素比特征向量更接近人们对图象内容的认知,更为直观,并具有一定的视觉意义,它可以看作是处于底层特征向量与高层语义之间的一个中间层.

如前所述,WWW 上存在着海量的图象.搜索引擎从网上搜集图象并建立图象库供用户检索.然而浏览这样大规模的图象库,对于用户来说是不可能的.对图象库中的图象进行分类使得对大规模图象库的访问更为有效,另外,分类通过剔除

收稿日期:2002-03-20;修回日期:2002-06-04
基金项目:国家自然科学基金(No.60172025)

不相关类别中的图象缩小了检索空间[4].一般来说,基于特征向量的图象分类算法通常从以下的思想出发[5,6]:每幅图象与特征空间中的一个点即特征向量相对应,图象之间的相似度由特征向量之间的距离来度量.已标注的图象作为所属类的样本,对每个未标注的图象计算其与所有样本之间的距离,将它分到距离最近的样本图象所属的类中.整个过程中有两个关键问题[5]:第一,采用什么特征向量表示图象;第二,采用什么样的距离度量方法.然而基于特征向量的图象分类方法计算出的图象之间的距离有时在视觉上很不直观,例如人们很难判断一幅人物的图象与一幅猫的图象之间的距离和一幅人物的图象与一幅花的图象之间的距离哪个大哪个小.本文提出的基于特征元素的图象分类方法不以图象之间的距离作为分类的依据,而是试图从训练集中发现图象的特征元素与图象所属类别之间的关联规则,将这些规则用于未标注图象的分类.

2　特征元素的提取

在进行分类之前,首先提取所有图象的特征元素.本文采用了两种特征元素:颜色特征元素和形状特征元素.

2.1　颜色特征元素

文献[7]提出了特征元素的概念和一种颜色特征元素的表示方法.本文中的颜色特征元素就是在它的基础上构成的.首先,对图象进行颜色聚类,即将视觉上颜色相似的象素合并,这样图象就被分割为几个颜色区域.具体方法是在 HSV 颜色空间中对图象色度直方图进行无监督聚类,将直方图中相邻,即视觉上相似的颜色合并,得到几个颜色聚类区(具体使用了客观的 K 均值聚类算法).对一幅图象,所得到的聚类区的数目取决于图象中颜色的丰富程度,颜色比较丰富则类数会多些,反之则少些.然后对于每个颜色聚类区,计算以下参数来表示该聚类区的特性:

(1) AC:颜色聚类区在色度直方图上的中心点,表示该聚类区的色度信息;

(2) $center$:聚类区质心在图象中的几何位置;

(3) $total$:聚类区包含的象素总数,即聚类区的几何大小;

(4) sd:定义 $sd = 0.95 * \frac{total}{Area(0.95)}$,其中 $Area(0.95)$ 为能够包含聚类区中 95% 象素的最小的长方形所包含的象素总数,显然 sd 表示了颜色聚类区的空间和形状信息;

(5)最后,对每个颜色聚类区还计算了 CCV 和 CAC 特征,它们同样用来表示颜色聚类区的各种空间信息,包括离散程度等等.具体的计算方法可参见文献[7].

以上这些参数均为数值,而不是向量.除 AC 本身为离散的之外,其余参数还需要进行简单的离散化,它们都是颜色特征元素的组件.简单地说,颜色特征元素的提取首先将图象分割为几个颜色聚类区,然后分别计算各个聚类区的参数,用这些参数构成颜色特征元素.这样得到的颜色特征元素,其组件全部为离散的数值,并且每个组件都具有比较直观的视觉意义.与向量空间中的连续的特征向量相比,这样得到的特征元素更为贴近人们对图象的主观认知.

2.2　形状特征元素

与颜色特征元素不同,形状特征元素不是从图象直接计算出来的,而是通过形状特征向量来构建的.本文所采用的形状特征提取方法基于小波模极大值和不变矩[8].从视觉角度来看,图象的小波变换模的极大值点位于图象的边界上,也就是说,小波变换的结果可以指示出图象的边界.而不变矩具有平移、尺度、旋转不变性,所以采用七个不变矩来表示小波变换后的多尺度边界图象的形状特征.形状特征向量的提取依照下列步骤进行:

(1)对图象进行小波分解,得到多尺度的模图象;

(2)记录下小波变换域中模是局部极大值并且模大于事先设定的阈值的那些点,得到多尺度的边界图象;

(3)对每一尺度的边界图象计算出它的七个不变矩,所有尺度上的不变矩共同组成这个图象的特征向量;

(4)对特征向量进行归一化;

实验结果表明小波模极大值在分解层数大于 6 以上时几乎就没有什么区分能力了,因此小波分解的层数设为 6 就足够了.这样对图象进行 6 层小波分解并在每个尺度的边界图象上分别计算 7 个不变矩就可得到一个 42 维的向量.为构建特征元素可用某种方法对这些特征向量进行离散化.一个最直接的方法就是对特征向量进行聚类,聚类之后,每个特征向量属于一个聚类区,可将这个对应的聚类区作为特征元素的组件.但是如果仅用整个特征向量进行聚类,然后将对应的聚类区作为形状特征元素唯一的组件会给后续的处理带来困难,因为仅用一个聚类区来表示图象的形状信息是不够的.因此我们将整个特征向量分裂为几个子向量用于聚类.如前所述,整个形状特征向量是 42 维的,由 6 层,每层 7 个不变矩构成.因此我们将整个特征向量分裂为 6 个子向量,每个子向量是 7 维,表示一层上的 7 个不变矩.另外再将整个特征向量分裂为 7 个子向量,每个子向量是 6 维,表示 6 层上的同一个不变矩.形状特征向量的分裂过程如图 1.将所有的子向量都用于聚类,并将它们对应的聚类区作为形状特征元素的组件.这样加上整个特征向量对应的聚类区,形状特征元素总共包含 14 个组件.

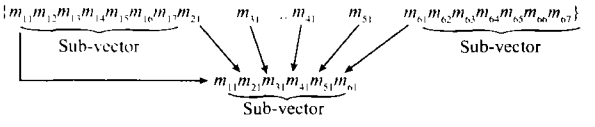

图 1　形状特征向量分裂过程

3　基于特征元素的图象分类

基于特征向量的图象分类算法通过计算特征向量之间的距离进行分类,但这种方法不适用于离散的特征元素.另一方面,定义特征元素之间的距离是非常困难的,因为特征元素之间的距离不具有直观的意义.所以在特征元素的基础上进行图象分类需要研究新的方法.本文提出通过关联规则挖掘来进行基于特征元素的图象分类.即用关联规则挖掘的方法发现图象的特征元素和图象所属类别之间的规则,然后运用这些规则预测未标注图象所属的类别.与基于特征向量的图象

分类算法相比,这种方法与人们认知图象的方式更为相近.

3.1 基于关联规则的分类

关联规则挖掘是目前在知识发现和数据挖掘领域最受关注的模式发现方法之一[9].简单地说,关联规则就是指形如 $X \Rightarrow Y$ 的表达式,其中 X 和 Y 是项目集.这种规则具有非常直观的含义:给定一个事务数据库 D,对于每条事务 $T \in D$,$X \Rightarrow Y$ 表示每当有一条事务 T 包含 X,那么 T 很可能也包含 Y.关联规则的支持率定义为联合概率 $p(X \subseteq T, Y \subseteq T)$,置信度定义为条件概率 $p(Y \subseteq T | X \subseteq T)$.如果一条关联规则的支持率和置信度分别大于指定的最小支持率和最小置信度,则称这条关联规则为强关联规则.关联规则挖掘就是要找出所有存在于事务数据库 D 中的强关联规则.

基于关联规则的分类(CBA)[10]的主导思想就是将关联规则挖掘用于分类.显然如果将关联规则的右边限定为类别,那么就可以将这种关联规则看作分类规则并用来建立分类器.数据集可以看作是一个包含 N 个数据项的关系表,每个数据项由 L 个独立的属性描述.这些数据项分别属于 Q 个已知的类别.对于离散的属性值,所有可能的值可直接被映射为一组正整数.对于连续的属性值,可将取值区间先离散化,而后同样映射为正整数.这里将属性值与正整数相映射是为了挖掘规则更方便所采取的一种手段.这样映射后,一个数据项就可以看作是一组(属性,整数)对和一个包含类别属性的项目集.每个(属性,整数)对为一个项目.顺便指出,图象特征元素的每个组件,即每个属性都是独立的,处于同等重要的地位.对每个属性进行的处理是相同的.

现在设 D 为数据集,I 为 D 中所有项目的集合,Y 为类别属性的集合.则表达式 $X \Rightarrow y$ 称为分类关联规则,其中 $X \subseteq I$ 且 $y \in Y$.一个数据项 $d \in D$ 表示 d 为一个项目子集,即 $X \subseteq d$ 且 $X \subseteq I$.如果 D 中 z 所有包含 X 的数据项中有 $c\%$ 标注为类别 y,则分类关联规则 $X \Rightarrow y$ 的置信度为 c.如果 D 中有 $s\%$ 的数据项包含 X 并且标注为类别 y,则分类关联规则 $X \subseteq y$ 的支持率为 s.CBA 的目的就是找到所有能够满足最小置信度和最小支持率限制的分类关联规则.对于标准数据集的实验结果表明 CBA 的分类准确度高于大数据集常用分类算法 C4.5[9].

3.2 算法过程

首先,按照第 2 节中所述的方法对所有图象提取特征元素.将每个图象看作一个数据项,特征元素的每个组件即为数据项的属性,也就是项目.训练图象集中的每个数据项均含有类别属性.与 CBA 中的初始假设略有不同的是,由于颜色特征元素提取过程中,不同图象颜色聚类区的数目不固定,每个图象所包含的项目数目不是固定数值.特征元素提取完成后,可以对训练图象集形成一个关系表.我们运用 CBA 对训练图象集所形成的关系表进行分类关联规则挖掘,挖掘得到的规则的数目取决于训练集的大小,训练集越大,规则挖掘所需的时间越长.得到的分类关联规则将用于建立图象分类器,即根据这些规则预测未标注图象的类别属性.当一个未标注图象满足多个分类关联规则并得到不同的预测结果时,分类器将根据分类关联规则的支持率和置信度选择最终的结果.

4 实验结果

为了与基于特征向量的图象分类算法相比较,我们实现了一种最近特征直线算法(NFL)[5].该算法着眼于距离度量方法的改进.同其它基于特征向量的图象分类算法相同,每幅图象对应于一个特征向量,即特征空间中的一个点.特征空间中连接两个点的直线称为特征直线.每类中所有样本图象,即已标注图象两两之间均连接一条特征直线.每个未标注的图象与某一类的距离就是它在特征空间中对应的点到该类所有特征直线的距离中最小的一个.分类的结果就是将未标注的图象分到与它距离最小的类别中.我们实现的 NFL 算法采用了颜色特征向量 *CCV*、*CAC* 以及基于小波和矩的形状特征向量[8].

我们使用的实验图象集含有 2558 幅图象,可分为 5 类,分别为汽车:505 幅、人像:565 幅、花朵:485 幅、风景:500 幅和花丛:503 幅.其中汽车、人像和花朵类的图象具有比较明显的目标,花丛和风景类的图象没有明显的目标或者目标比较分散.风景类图象包括日出、日落、海滩、山脉、森林等等.特征元素和特征向量提取完成之后,整个图象集被划分为测试集和训练集.我们进行了两次实验,每次实验均从整个图象集中取 1/3 作为测试集,余下 2/3 作为训练集,两次实验所取的测试集没有交集,即测试集的图象完全不同.分别用两种方法对两个测试集进行分类所得到的错误概率如表 1.

表 1 两种方法分类错误率比较

错误概率	测试集 1		测试集 2	
	FEBIC	NFL	FEBIC	NFL
汽车	21.3%	23.1%	18.3%	23.1%
人像	22.9%	25.6%	0.7%	26.1%
花朵	32.1%	48.8%	36.4%	46.9%
风景	30.7%	38.0%	32.5%	34.3%
花丛	26.8%	45.8%	20.2%	37.0%
总计	26.6%	35.8%	25.4%	33.2%

可以看出基于特征元素的分类方法错误率较低.另外,对于相同大小的训练集和测试集,基于特征元素的分类方法的时间复杂度大大低于 NFL.因为 NFL 需要进行大量的加减乘除和比较大小的数学运算,而基于特征元素的图象分类算法所需的数学运算很少.以上面的训练集和测试集为例,NFL 计算一个未标注的图象到各类的距离并找出最小的距离需时 18 秒左右,整个测试集需时 4 至 5 小时.而基于特征元素的分类方法运用 CBA 挖掘分类关联规则需时 40 秒左右,应用分类规则预测测试集图象的类别需时 20 秒左右,整个过程仅需时 1 分钟左右.

5 结论与展望

本文提出了一种基于特征元素的图象分类方法.与特征向量相比,特征元素能够根据人们对图象的主观认知反映图象的视觉内容.在基于特征元素的图象分类算法中,不是计算图象与图象之间的距离,而是通过发现图象的特征元素与图象所属类别之间的关联规则建立分类器.实验结果表明,基于

特征元素的分类方法与基于特征向量的分类方法相比能够达到更小的错误率.而且它还可大大缩短计算时间,更适用于大规模的图象库.目前我们仅仅完成了一些初步的工作,其中形状特征元素仅在特征向量的基础上进行了离散化,仍未完全摆脱特征向量不具有直观视觉意义的弱点,需要进行进一步的研究和改进.另外,我们还将研究纹理等其它特征元素的提取方法,从而进一步提高分类的准确率.

参考文献:

[1] S Mukherjea, J H Cho. Automatically determining semantics for world wide web multimedia information retrieval [J]. Journal of Visual Languages and Computing, 1999, 585 - 606.

[2] J R Smith, S-F Chang. Visually searching the web for content [J]. IEEE Multimedia Magazine, 1997, 12 - 20.

[3] Y Xu, Y-J Zhang. Feature element theory for image recognition and retrieval [A]. Proc Storage and Retrieval for Media Databases [C]. Bellingham: SPIE, 4676, 2002. 126 - 137.

[4] K Hirata, S Mukherjea, W-S Li, Y Hara. Integration of image matching and classification for multimedia navigation [J]. Multimedia Tools and Applications, 2000, 11: 295 - 309.

[5] S Z Li, K L Chan, C-L Wang. Performance evaluation of the nearest feature line method in image classification and retrieval [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2000, 22(11): 1335 - 1339.

[6] A Mittal, L-F Cheong. Techniques for designing a classifier for multimedia indexing [A]. Proc Storage and Retrieval for Media Databases [C]. Bellingham: SPIE, 4315, 2001. 107 - 117.

[7] Y Xu, Y-J Zhang. Image retrieval framework Driven by Association Feedback with feature elements evaluation built in [A]. Proc Storage and Retrieval for Media Databases [C]. Bellingham: SPIE, 4315, 2001. 118 - 129.

[8] Y-R Yao, Y-J Zhang. Shape-based image retrieval using wavelets and moments [A]. Proc of Workshop on Very Low Bitrate Video'99, Kyoto [C]. Japan: JSPS 159th Committee, 1999. 71 - 74.

[9] J Hipp, U Guntzer, G Nakhaeizadeh. algorithms for association rule mining - a general survey and comparison [J]. ACM SIGKDD Explorations Newsletter, 2000, 2(1): 58 - 64.

[10] B Liu, W Hsu, Y-M Ma. Integrating classification and association rule mining [A]. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD - 98) [C]. New York USA: AAAI, 1998. 80 - 86.

作者简介:

李　勋　女,1977 年 5 月生于北京.1999 年 7 月于清华大学电子工程系获得学士学位.2002 年 7 月获清华大学电子工程系硕士学位.研究方向包括多媒体数据库以及基于网络的图象分类和图象检索等.Email: liq@image.ee.tsinghua.edu.cn

章毓晋　男,1954 年生于太原.清华大学电子工程系图象图形研究所副所长,教授,博士生导师.IEEE 高级会员,《中国图象图形学报》副主编,"Pattern Recognition Letters","International Journal of Image and Graphics",《电子与信息学报》,《计算机辅助设计与图形学学报》编委,第一届和第二届国际图象图形学术大会(ICIG)程序委员会主席.主要研究领域是图象工程(图象处理,图象分析,图象理解及其技术应用),已发表 180 多篇研究论文,著有《图象分割》等书 5 本.Email: zhangyj@image.ee.tsinghua.edu.cn