基于 Boosting 学习的图片自动语义标注

茹立云 马少平 路 晶
(清华大学计算机科学与技术系 智能技术与系统国家重点实验室, 北京 100084)

摘 要 图片自动语义标注是基于内容图像检索中很重要且很有挑战性的工作。本文提出了一种基于 Boosting 学习的图片自动语义标注方法,建立了一个图片语义标注系统 BLIR(boosting for linguistic indexing image retrieval system)。假设一组具有同一语义的图像能够用一个由一组特征组合而成的视觉模型来表示。2D-MHMM(2 维多分辨率隐马尔科夫模型)实际上就是一种颜色和纹理特殊组合的模板。BLIR 系统首先生成大量的 2D-MHMM 模型,然后用 Boosting 算法来实现关键词与 2D-MHMM 模型的关联。在一个包含60 000张图像的图库上实现并测试了这个系统。结果表明,对这些测试图像,BLIR 方法比其他方法具有更高的检索正确率。

关键词 基于内容图像检索 图像语义标注 Boosting 算法 2 维多分辨率隐马尔科夫模型(2D-MHMM)
中图法分类号: TP37 **文献标识码**: A **文章编号**: 1006-8961(2006)04-0486-06

Boosting-based Automatic Linguistic Indexing of Pictures

RU Li-yun, MA Shao-ping, LU Jing
(*The State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Automatic linguistic indexing of pictures is an important but highly challenging problem for researchers in content-based image retrieval. In this paper, a boosting-based automatic linguistic indexing approach is proposed and a linguistic indexing system called BLIR(Boosting for Linguistic indexing Image Retrieval system) is built. It is assumed that images of same semantic meaning can be represented by a model combined with a group of features. 2D-MHMM model is found to be such a template for one special kind of color and texture combination, which corresponds to one cluster in feature space. Thus in BLIR system, a large number of 2D-MHMM models are generated and a boosting algorithm is used to associate keywords with models. The system has been implemented and tested on a photographic image database of about 60 000 images. Results demonstrate the effectiveness of the proposed technique which outperforms other approaches.

Keywords content-based image retrieval, linguistic indexing of pictures, Bootsting algorithm, two-dimensional multi-resolution hidden Markov model(2D-MHMM)

1 引 言

图像检索的根本问题是一个视觉问题,即让计算机基于语义来理解数据库中的图片,例如一张包含人物的图片,计算机要把图片中的人物、位置以及其他物体用语言文字来表述出来。如果能做到这一步,那么现有的图像检索问题实际上就可以转化成技术已经相当成熟的文本检索问题。图片自动语义标注在基于内容图像检索和计算机物体识别中是相当重要的。它的潜在应用领域包括生物医学、商业、军事、教育、数字图书馆和网上检索等。

然而这个视觉问题在现阶段还是不可解的,因为现阶段自然语言理解、图片理解都尚未达到能够

基金项目:国家重点基础研究发展计划("973")项目(2004CB318108);国家自然科学基金项目(60223004, 60321002, 60303005);教育部科学技术研究重大项目(104236)
收稿日期:2004-11-09;**改回日期**:2005-06-22
第一作者简介:茹立云(1979 ~),男。2005 年于清华大学计算机科学与技术系获硕士学位。主要研究方向为信息检索、机器学习等。E-mail:lyru98@mails.tsinghua.edu.cn

实用的地步。但是,图片的很多语义都是和一些颜色,或纹理,或形状的特征相关的,把这些特征的组合称为视觉特征模型。

因此如果能够自动将与某个语义特征相对应的视觉特征模型找出来,那么图像检索就变成了利用和某个语义特征关联的模型而进行的检索。这样做将会提高基于内容的图像检索的正确率。本文提出了一种基于 Boosting 学习的图片自动语义标注方法,以此为基础可以得到一种基于内容图像检索系统的架构。基本思想是:首先构造很多的模型,然后在模型和概念之间建立联系,保持一种多对多关系。一个 2D-MHMM(2 维多分辨率隐马尔科夫模型)模型可以被看作是特征空间中的一个聚类,这样就能产生许多 2D-MHMM 模型,然后再用 Boosting 算法将概念与 2D-MHMM 模型建立连接。在对每个图片进行了语义标注之后,就能以关键词的方式进行检索。

2　相关工作

基于内容的图像检索是这些年很热门的研究领域。自 20 世纪 90 年代初期以来,研究者已经开发了许多基于内容的图像检索系统[1~4]。其中的大部分系统用诸如颜色、纹理、形状等特征来表示图像,检索系统主要是检索与查询图像或检索草图视觉相近的图像。然而由于图像底层特征与高层语义之间的不一致性,且由于对大量物体的识别存在很大的困难,因此这些系统一般都不具有自动给图片分配易理解的文本描述(如语义标注)的能力。然而,这个功能对于将图片和文本结合起来是很重要的,并且它能拓宽图片库可能的应用。

将图像跟单词自动关联起来是弥补上述不一致性的一个可能的解决办法。基于学习的语义标注系统首先用大量经过标注的图片来训练,再用这些经过训练的模型来标注新的图片。Minka[5] 提出了一种基于多个模型的图片理解框架,该系统能帮助用户标出某个概念所在的区域。这个系统对于单幅图片的标注效果很好,但对于图片间的学习和标注的扩展性不是很强。

Barnard 和 Forsyth 使用图像分割的特征来学习图片的语义[6]。但是,由于这种做法是建立在对图片正确分割的基础上的,而图片自动分割仍然是计算机视觉领域的一个开放性问题[7,8]。特别对于区域特征不明显的图片,例如人物、建筑等,这种做法的效果往往不是很好。

Li 和 Wang 提出了一种基于统计建模方法的图片自动语义标注方法[9]。他们用了一种在计算机视觉中用于图片分类的 2 维多分辨率隐马尔科夫模型(2D-MHMM)[10],这种模型的优势在于它可以对任一组图片建立一个统计模型。该统计模型相当于一种特定的纹理。这个方法在具有特定概念的图片上具有很好的效果。

Minka 和 Li 的方法都是假设文字描述的概念可以用一种模型来表达的,而事实上,这种假定往往是不可靠的,因为对于一些比较复杂的概念很难用一种模型来描述。如图 1 所示是在 Corel 的图片库中人物(people)概念的表述,可以看出,这种概念是很难用一个特征组合,或者某种纹理来表达的。

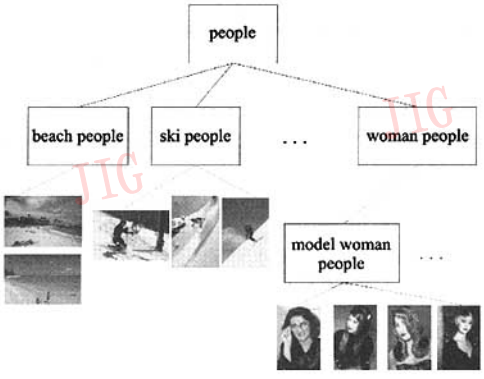

图 1　一些概念很难用单个模型来表述

Fig. 1　Hard to represent some concept with a single model

3　2 维多分辨率隐马尔科夫模型

2 维多分辨率隐马尔科夫模型(2D-MHMM)是一个统计图像建模方法[10]。它试图用统计的方法从训练图像中学习一种“纹理”。它用一个多层(多分辨率)2 维隐马尔科夫模型(2D-HMM)来对图像建立模型。不同分辨率之间的关系就像一棵四叉树,如图 2 所示。

在每一种分辨率下,用一个 2 维隐马尔科夫模型[11]对图像建立模型。图像被分成块,假设每个块只依赖于它上方和左方的块。对每个块抽取一个特征向量,2D-HMM 的状态也是特征向量,对每个状态,假设其特征向量满足高斯分布。状态通过 EM(expectation maximization)算法来计算,然后用这些状态来训练 2D-HMM。概率用 Viterbi 算法来计算。

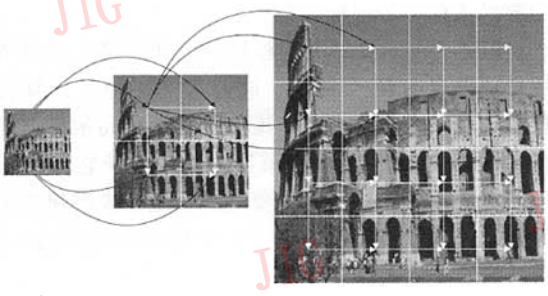

图 2 2D-MHMM 模型中块之间的空间依赖性

Fig. 2 The spatial dependency between blocks in 2D-MHMM

对于多分辨率隐马尔科夫模型(MHMM),用 $\boldsymbol{R}=\{1,\cdots,R\}$ 来表示分辨率集合,其中 $r=R$ 表示最精细的分辨率。设分辨率 r 下块的标记集合为 $\boldsymbol{N}^{(r)}=\{(i,j):0\leqslant i<w/2^{R-r},0\leqslant j<z/2^{R-r}\}$。图像用所有分辨率下的特征向量来描述,表示为 $\boldsymbol{u}_{i,j}^{(r)}$, $r\in\boldsymbol{R}$。特征向量的潜在状态是 $s_{i,j}^{(r)}$。在每个分辨率 r 下,状态集是 $\{1^{(r)},2^{(r)},\cdots,S_r^{(r)}\}$。

为了构建分辨率间的统计依赖性,2D-MHMM 中假设了一个由分辨率来扮演时间角色的马尔科夫链。每个块只依赖于它的父块,这样给定父分辨率下的状态和特征,当前分辨率下的状态和特征就跟其他先前的分辨率条件独立,所以

$$P\{s_{i,j}^{(r)},\boldsymbol{u}_{i,j}^{(r)}:r\in\boldsymbol{R},(i,j)\in\boldsymbol{N}^{(r)}\}= \tag{1}$$

$$P\{s_{i,j}^{(1)},\boldsymbol{u}_{i,j}^{(1)}:(i,j)\in\boldsymbol{N}^{(1)}\}\times$$

$$P\{s_{i,j}^{(2)},\boldsymbol{u}_{i,j}^{(2)}:(i,j)\in\boldsymbol{N}^{(2)}\mid s_{k,l}^{(1)}:(k,l)\in\boldsymbol{N}^{(1)}\}\times\cdots\times$$

$$P\{s_{i,j}^{(R)},\boldsymbol{u}_{i,j}^{(R)}:(i,j)\in\boldsymbol{N}^{(R)}\mid s_{k,l}^{(R-1)}:(k,l)\in\boldsymbol{N}^{(R-1)}\} \tag{2}$$

在最粗糙的分辨率 $r=1$ 下,假设特征向量由一个单分辨率的 2D-HMM 产生。在更高一层的分辨率下,假设给定状态下的特征向量的条件分布是一个高斯分布。高斯分布的参数依赖于特定分辨率下的状态。

给定分辨率 $r-1$ 下的状态,在更好分辨率 r 下,块之间的统计依赖性受兄弟块(从同一父块传下来的子块)的约束。明确地说,从不同父块传下来的子块是条件独立的。此外,给定父块的状态,它的子块状态独立于它们的"叔伯"块(父分辨率下的非父块)的状态。兄弟块之间的状态转换由马尔科夫链的特性支配,这个特性是跟单分辨率下的 2D-HMM 的假设一样的。然而,状态转移概率依赖于它们父块的状态。用公式来表示这些假设,分辨率 $r-1$ 的块 (k,l) 在分辨率 r 下的子块用 $\boldsymbol{D}(k,l)=\{(2k,2l),(2k+1,2l),(2k,2l+1),(2k+1,2l+1)\}$ 表示,根据假设

$$P\{s_{i,j}^{(r)}:(i,j)\in\boldsymbol{N}^{(r)}\mid s_{k,l}^{(r-1)}:(k,l)\in\boldsymbol{N}^{(r-1)}\}=$$

$$\prod_{(k,l)\in\boldsymbol{N}^{(r-1)}}P\{s_{i,j}^{(r)}:(i,j)\in\boldsymbol{D}(k,l)\mid s_{k,l}^{(r-1)}\}$$

其中,$P\{s_{i,j}^{(r)}:(i,j)\in\boldsymbol{D}(k,l)\mid s_{k,l}^{(r-1)}\}$ 可以由在条件 $s_{k,l}^{(r-1)}$ 上的转移概率来计算,表示为 $a_{m,n,l}(s_{k,l}^{(r-1)})$。这样就有了对父分辨率下每个可能状态的一组不同的转移概率 $a_{m,n,l}$。先前分辨率的影响通过状态的概率分层地施加。然后式(2)中在所有分辨率下状态和特征向量的联合概率被导出。

4 基于 Boosting 学习的图片自动语义标注

4.1 Boosting 算法

机器学习的分类器算法按分类能力分可分为弱分类器和强分类器两大类。Boosting 算法是为了解决把弱分类器的分类精度提高到和强分类器相匹配的问题而提出的。

Boosting 算法的基本思想是:

(1)每个样本都赋予一个权重;

(2)T 次迭代,每次迭代后,对分类错误的样本加大权重,使得下一次的迭代更加关注这些样本。

Ada-Boosting[12] 是一种常用的 Boosting 算法,Ada 指 Adaptive,即这种 Boosting 算法具有较强的适应性。Ada-Boosting 的主要思想是保持一个带有分布特征的训练数据集,每一次迭代时,都调整该数据集的分布特征,从而产生新的分类器。刚开始时,训练数据初始权重是相同的。每一次迭代,算法增加错分类的数据的权重,降低正确分类的数据权重,这样使新的分类器重点放在那些分类困难的数据上。最终的分类器由若干弱分类器加权而成,弱分类器对训练数据的分类能力越强则权重越高。

4.2 用 Ada-Boosting 实现模型与关键词的关联

假设已经产生了很多的模型,对于区分某个关键词,即是否在某张图片上标注某个关键词,假定在某个模型下仅仅设定一个门限值就可以将两者分开。之所以选择 Ada-Boosting 算法,是因为(1)由于有的语义概念非常复杂,在每个模型上区分是否包含某个语义概念相当于一个弱的分类器;(2)Ada-Boosting 需要的样本数不是很大。(由于语义概念很多,因此具有某个语义概念的样本数不会很多,可以对小样本进行训练是一个重要的条件)

用数据库中所有标注的关键词作为训练数据。

基本思想是:每次从模型中挑出最能在训练集上区分是否应该标注关键词的模型,然后利用 Boosting 的方法调整训练数据的权重。以此循环下去,形成模型和关键词的关联。算法如下:

输入:$(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$

$Y_i \in \{-1,+1\}$ 代表是否标有某个关键词

初始化:$D_1(i)=\frac{1}{n}$,n 为训练数据个数

for $t=1,\cdots,T$

在 D_t 下,

对于任意模型 M_i,训练得门限值 f_{it},得到假设

$$h_{it}(X)=\begin{cases}+1 & M_i(X)>f_{it}\\ -1 & M_i(X)\leqslant f_{it}\end{cases}$$

计算错误率 $E_{it}=\sum_{i(h(X_i)\neq Y_i)} D_t(i)$

取最佳的模型 M_k,$\forall i, E_{kt}<E_{it}$

得到弱的假设 $h_t=h_{kt}: Y\in\{-1,+1\}$

错误率:$E_t=\sum_{i(h(X_i)\neq Y_i)} D_t(i)$

选择 $\alpha_t=\frac{1}{2}\ln\left(\frac{1-E_t}{E_t}\right)$

更改权值:

if　$h_t(X_i)\neq Y_i, D_{t+1}(i)=D_t\cdot e^{\alpha_t/Z_t}$

else　$h_t(X_i)=Y_i, D_{t+1}(i)=D_t\cdot e^{-\alpha_t/Z_t}$

其中,Z_t 为一个归一化分布的值。

End for

输出:$H(X)=\sum_t \alpha_t h_t(X)$　　(3)

图 3 是用以上算法生成的框架。

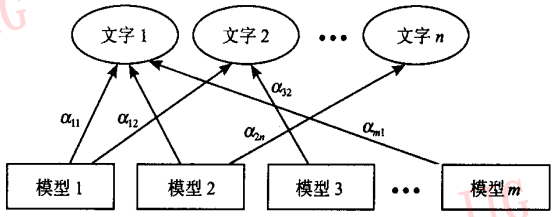

图 3　Ada-Boosting 生成的框架

Fig. 3　Architecture generated by Ada-Boosting

由图 3 可以看出,该算法在模型和文字之间通过式(3)建立了对应关系,每一个连接有一个权值 α_t,反映了文字和某个模型联系的强弱,$H(X)$ 表示图像 X 标注某个关键词的可能性,值越大表示标注该关键词的可能性越大,反之则越小。

图片语义标注系统 BLIR 的基本思想是:首先生成大量的和语义高度相关的视觉模型,然后利用图 3 的学习方法把语义概念和视觉模型联系起来,再以关键词的方式进行检索,检索结果按图片标注该关键词可能性的大小进行排序。整个系统架构如图 4 所示。

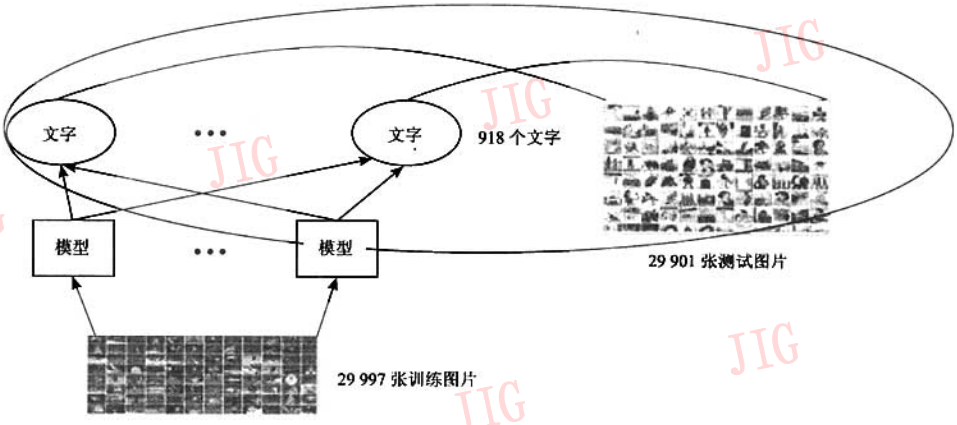

图 4　BLIR 图像语义标注系统架构

Fig. 4　Architecture of BLIR system

5　实验与分析

5.1　实验设置

用 Corel 图库中 600 个 CD-ROMs 的60 000张图片作为实验数据以证明本文方法的有效性,其中每个 CD-ROMs 包含 100 张图片,代表一组图片。每张图片都对应地标注有 4 个左右的关键词。把数据分成训练集和测试集,首先在训练集上得到每组图片的模型,然后对训练集进行训练得到模型和关键词

的对应关系,这样利用训练后的这种对应关系就可以对测试集中的图片进行标注。利用标注后的结果可以对某个语义概念在数据库中进行检索,并评估正确率。检索结果排序按照式(3)的结果。

实验主要包括两部分,具有相关主题的图像库上的测试以及独立主题的图像库上的测试。所谓相关主题的测试库是指测试集中的图片跟训练集来自于 Corel 图库同一组 CD-ROMs,也就是说把一组 CD-ROMs 的一些图片用作训练,而剩余的那些图片用作测试;而所谓完全独立主题的测试库是指测试图像来自于跟训练集没有交集的另一组 CD-ROMs。

5.2 实验结果

从 600 个 CD-ROMs 中随机选择其中的 300 个 CD-ROMs,并从每个 CD-ROM 中随机选择 60 张图片作为训练数据,然后将每个 CD-ROM 中剩余的 40 张图片作为第 1 个实验的测试数据,这样就得到一个拥有 300 组,每组包含 60 张,共计18 000张图片的训练集。首先对每个 CD-ROM 中的训练数据生成一个 2D-MHMM 模型,生成一个概念词典,然后统计训练数据中出现的关键词,形成一个关键词词典;再用 Boosting 学习算法将两个词典关联起来。在实验中从训练数据统计得到一个包含 918 个关键词的关键词词典。然后根据式(3)可以得到图像标注关键词词典中任一关键词的可能性大小,在检索某特定关键词的图像时可以通过该可能性的大小来排序。对于每一个语义关键词,通过看检索结果中具有该语义的图片个数来计算检索的正确率。

由于 Corel 图库的每个 CD-ROM 内图像具有相似性,该实验主要验证 BLIR 系统架构的可扩展性。统计对于每个关键词的前 100 个检索结果的正确率,并将本文算法和 Li[9] 的 ALIP 系统的算法进行了比较,由于 ALIP 系统的算法和 BLIR 系统采用了同样的视觉模型集合,而 ALIP 系统的方法在模型集合上用了贝叶斯的方法,这个实验主要说明本文机器学习方法的效果。

实验结果如图 5 和图 6 所示,其中,图 5 显示了在训练集上关键词的前 100 个检索结果的平均正确率,图 6 显示了在测试集上关键词的前 100 个检索结果的平均正确率。其中,BLIR 方法是指本文 BLIR 系统的方法,ALIP 方法是指 Li 和 Wang 的 ALIP 系统采用的方法。从实验结果可以看出本文方法是相当有效的。

第 2 个实验是在完全独立主题的测试库上进行的。从 600 个 CD-ROMs 中随机选择其中的 300 个 CD-ROMs,共 30 000 张图片作为训练集,每个 CD-ROM对应地生成一个 2D-MHMM 模型,用另外的 300 个 CD-ROMs 的30 000张图片作为测试集。这两个集合没有交叉。同样进行如上的实验,实验结果如图 7 所示。

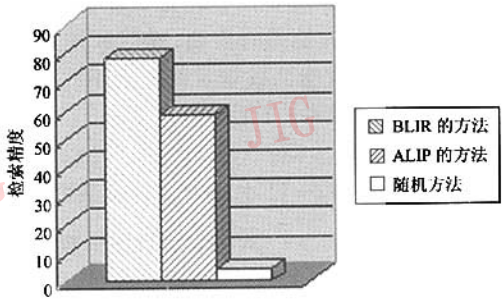

图 5 在训练集上关键词的前 100 个检索结果正确率

Fig. 5 Accuracy after the first 100 retrieved results for the keywords in the training set

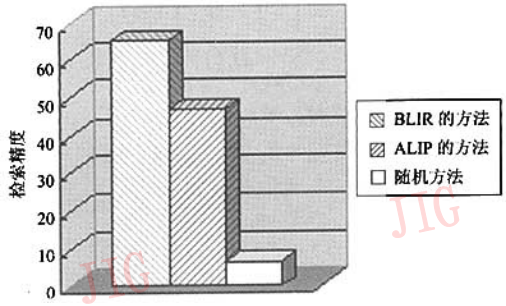

图 6 在测试集上关键词的前 100 个检索结果正确率

Fig. 6 Accuracy after the first 100 retrieved results for the keywords in the test set

从图 7 中可以看出,没有主题重复的测试结果比有主题重复的结果差些。这表明了本文算法还是

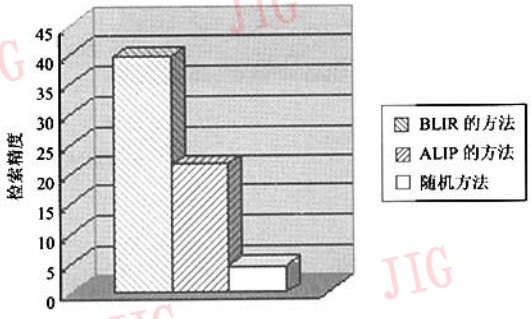

图 7 没有主题重复测试集上的实验结果

Fig. 7 Results for the irrelevant test set

依赖于描述相同语义内容是视觉相似的这个假设的。从图 7 还可以看出,本文算法依然得到了最好的结果。

6 结　论

基于学习的语义标注系统首先用大量经过标注的图片来训练,再用这些经过训练的模型来标注新的图片。将图像跟单词自动关联起来是弥补图像高层语义含义与低层特征之间不一致性的一个可能的解决办法。本文提出了一种可以用于图像自动语义标注的架构。在 BLIR 系统中,用分类后的图像来自动训练数百个模型,然后用 boosting 学习的方法,在关键词和模型之间建立起一种多对多的联系,借助每个模型的响应来实现对关键词的自动标注。在 BLIR 系统中选择 2D-MHMM 作为模型支持。在一个具有60 000张图片的图库上进行了实验,并将该算法与 Li[9] 中的方法进行了比较,实验结果表明该算法能够获得更好的检索正确率,系统性能有 20% 左右的提高,从而证明了 BLIR 的系统架构是十分有效的。本文方法的主要优点有:(1)能够采用各种视觉特征模型,如 2D-MHMM, MRSAR 等;(2)它能通过 Adaboost 算法在模型和概念之间生成一个多对多的关系,从而能够表示复杂的概念;(3)它能够提供一个简化的图像检索界面,且具有很高的精度。

文中用单个模型来表示一组图像,但是对一些具有不同表现的复杂图像组,只用一个模型并不能有效地表示它。为了解决这个问题,将考虑采用自适应的方法来对图像组建立模型,即根据图像组的复杂性,系统会自动的判断需要多少个模型来表示它们。

参考文献(References)

1 Niblack W, Barber R, Equitz W, *et al.* The QBIC project: querying images by content using color, texture and shape [A]. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases[C], San Jose, CA, USA, 1993, **1908**:173 ~ 187.

2 Batch J R, Fuller C, Gupta A, *et al.* The virge image search engine: an open framework for image management[A]. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases[C], San Jose, CA, USA, 1996, **2670**:76 ~ 87.

3 Smith J R, Chang S F. An image and video search engine for the World-Wide Web[A]. In:Proceedings of SPIE [C], San Jose, CA, USA,1997, **3022**:84 ~ 95.

4 Wang J Z, Li J, Wiederhold G. SIMPLIcity: Semantics-sensitive integrated matching for pictures libraries[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, **23**(9): 947 ~ 963.

5 Minka TP, Picard R W. Interactive learning using a "society of models"[J]. Pattern Recognition, 1997, **30**(4):565 ~ 581.

6 Barnard K, Forsyth D. Learning the semantics of words and pictures [A]. In: Proceedings of International Conference on Computer Vision [C], Vancouver, Canada, 2001:408 ~ 415.

7 Zhu S, Yuille A L. Region competition: Unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, **18**(9): 884 ~ 900.

8 Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, **22**(8): 888 ~ 905.

9 Li J, Wang J Z. Automatic linguistic indexing of pictures by a statistical modeling approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, **25**(10):14.

10 Li J, Gray R M, Olshen R A. Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models [J]. IEEE Transactions on Information Theory, 2000, **46**(5): 1826 ~ 1841.

11 Li J, Najmi A, Gray R M. Image classification by a two-dimensional hidden Markov model[J]. IEEE Transactions on Signal Processing, 2000, **48**(2): 517 ~ 533.

12 Freund Y. An adaptive version of the boost by majority algorithm[J]. Machine Learning, 2001, **43**(3):293 ~ 318.