基于多证据融合的视频排序方法

韦世奎,赵 耀,朱振峰
(北京交通大学信息科学研究所,北京 100044)

摘 要: 在视频检索中,通过对用户行为特性的分析发现,用户通常只关注排在最前面的返回结果,而很少有耐心将所有的返回结果浏览一遍.因此,对于一个搜索引擎来说,能否将最相关的结果排在最前面是至关重要的.为了实现这一目标,本文提出了一种基于多证据融合的视频排序方法.该方法利用 Dempster-Shafer 证据推理理论来协同地融合多方证据,进而推断出最相关的视频镜头.如果多方证据一致,则证明某个视频镜头是相关的,此镜头被认为是最相关的镜头,并被排在返回列表的最前列.相反,如果多方证据产生冲突,那么此镜头就将被排在后面.实验结果表明,利用建议的多证据融合排序算法,搜索引擎的搜索质量,特别是排在前列的搜索结果的准确性,有了明显的改善.

关键词: 多证据融合; 视频排序; 检索; Dempster-Shafer 理论
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2010) 01-0167-06

Video Ranking with Multi-Evidence Combination

WEI Shi-kui, ZHAO Yao, ZHU Zhen-feng
(*Institute of Information Science, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract: According to the analysis on users' query behavior, we found that the users are rarely patient to go through all the returned results. Usually, they only check the most top returned results, so it is vital for a search engine to rank high relevant results to the top. In order to move vigorously toward this goal, we proposed a video ranking method based on multi-evidence combination. Its theoretical foundation is Dempster-Shafer evidence combination theory, which can fuse the evidence from distinct views and cooperatively infer the most likely relevant results. If the evidence from multiple aspects reaches a consensus on the conclusion that a certain video shot is relevant to query, the video shot is treated as the most likely relevant shot and is ranked to the top of the returned result list. On the contrary, a lack of consensus leads to the judgment that the video shot is not highly relevant to the query and is ranked to the bottom of the returned result list. Experimental results show that the search quality, especially on the most top-ranked results, is significantly improved after using the proposed scheme.

Key words: multi-evidence combination; video ranking; retrieval; Dempster-Shafer theory

1 引言

随着大量视频数据的涌现,视频搜索技术越来越显得重要.尽管各种各样的视频搜索模型被提出,但大部分模型致力于如何返回尽可能多、尽可能精确的结果.为此,研究者关注更多的是如何开发好的特征提取方案和有效的特征度量方法.然而,由于图像理解技术和视频描述技术的限制,低层特征相似性度量的结果通常和用户的期望有很大出入,这就是所谓的语义鸿沟问题[1].另外,此类方法也忽略了用户在实际查询过程中的行为特性.事实上,当用户搜索某一视频信息时,很少有耐心将搜索引擎返回的结果从头到尾浏览一遍.通常,他们只查看前几个结果,以判断是否满足自己的需要.因此,提高排在最前面结果的准确度对一个视频搜索引擎来说至关重要.尽管当前成功的视频搜索引擎取得了还算令人满意的查全率,但这些系统的准确率相对较低,特别是排在最前面结果中的正例较少[2].

作为一类有效的解决方案,重排序技术已经被成功地应用于网页搜索领域[3].它的目标是通过重新排列基于相似性度量的初始搜索结果,将真正相关的结果排在前面,即提高排在前面结果的准确度.尽管一些学者已经尝试将重排序的思想应用于视频检索领域,但这方面的工作还很少.作为一次尝试,文献[2]提出了一种基于信息瓶颈理论的排序技术,这种方法通过聚类得到降噪的后验概率,并为聚类后的每一类计算一个局部特征密度.利用降噪的后验概率和局部特征密度,可以为搜索结果列表中的每一个镜头计算一个分数,并依据此分数来重新排序镜头.文献[4]提出了一种基于模型的重排

收稿日期:2008-08-25;修回日期:2009-08-14
基金项目:国家自然科学基金(No.60602030,60776794);国家 863 高技术研究发展计划(No.2007AA01Z175);国家重点基础研究发展计划(No.2006CB303104);教育部长江学者和创新团队发展计划(No.IRT707);模式识别国家重点实验室开放基金

序技术.事实上,这种技术不同于传统意义上的重排序,因为它本身就可以作为一个独立的检索模型来使用.其基本思想是利用概念检测器来提高初始搜索的精度.和以上两个方法不同的是,文献[5]利用 Adaboost 算法从多个弱特征中学习一个新的检索模型来重新排列初始搜索结果.

尽管以上重排序方案在一定程度上提高了视频搜索的平均精度,但这些方案对于排在最前面的搜索结果给予了较少的关注,也就是说,所有的初始搜索结果在重排序过程中被同等的对待.从适应用户查询行为特性的角度来看,这些方案并不能很好地满足用户需求.为此,本文提出了一种基于多证据融合的排序算法.其基本思想是,利用 Dempster-Shafer 证据推理理论[6]来协同地融合多方证据,进而推断出最可能相关的镜头.和以往排序算法不同的是,本方法有区别地重新排列初始搜索镜头,从而给予排在最前列镜头更高的准确度.

2 多证据融合排序方案

在这一节,本文将详细地介绍建议的多证据融合排序方案.此方案主要包括三个方面,即定义识别帧(Frame of Discernment,简称 FOD)、设计信任函数、融合多个证据.此方案的整体框架结构如图 1 所示.注意,在算法的叙述和实验部分,本文从两个近似独立的证据空间(或特征空间)来观察证据,它可以很方便地推广到更多证据空间.下面详细介绍此算法的每一部分.

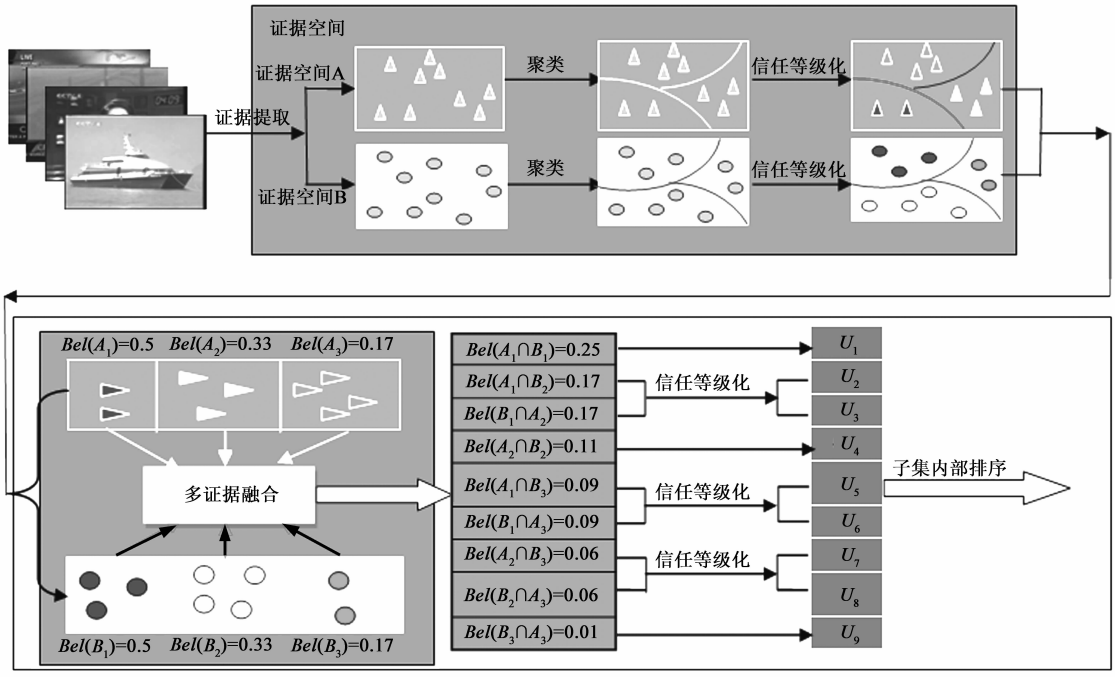

图1 多证据融合排序技术的整体系统框图

2.1 定义识别帧(FOD)

在 Dempster-Shafer 理论中,对一个特定的问题,通常对应一个互斥的完备答案集,这个答案集被称之为此问题的识别帧.例如在本应用中,我们的问题是:在初始搜索结果列表中,哪些镜头是真正和查询相关的.那么这个初始搜索列表就可以看作是此问题的一个识别帧,此识别帧中的元素就是镜头.对于其中任意一个元素(或镜头)E,可以作为此问题的一个回答,即 E 是最相关的.事实上,任何对这个初始搜索列表不重叠的划分,都可以形成一个互斥完备子集集合,并作为此问题的一个识别帧.在这种情况下,识别帧中的一个元素就是一个子集合.对于任意一个子集合 S,我们可以说最相关的镜头在子集合 S 中.

我们的最终目的是利用多个证据来协同推断出初始搜索列表中哪些镜头是最相关的.通过在不同证据空间对初始搜索结果进行不同划分,可以为同一个问题形成多个识别帧.这样,就可以通过为每一个识别帧设计一个信任函数来为识别帧的每一个元素分配一个信任值.融合多个信任函数,我们可以最终推断出那些镜头是最相关的.

在本算法中,识别帧的形成是通过在每一个证据空间将初始结果列表聚类为固定的几类来实现.这样,在每一个证据空间,定义了一个识别帧,它包括几个互斥的子集合.尽管我们的排序方案不依赖于特定的聚类算法,但聚类的性能确实间接地影响了排序质量.在我们聚类策略中,我们使用 NCuts 聚类算法.为了避免聚类过于零散,我们设定为三类.如图 1 所示,在两个证据空间 A 和 B 下,初始搜集结果都被聚为 3 类.值得注

意的是,一个帧元素不仅仅是一个镜头集合,而且还是一个断言,即断言最相关的镜头在此集合中.

2.2 设计信任函数

一旦我们在每一个证据空间获得了一个识别帧,那么对于帧中的任一答案(或断言),我们都需要判断其是否是可信的.也就是说,我们需要为每一个帧元素分配一个信任值.在本小节,我们将详细地介绍一种有效的信任值分配方法,即信任函数.

因为每一个帧元素都断言自己包含的镜头是最相关的,所以帧元素中的镜头和查询意图的相关程度决定了我们对帧元素的信任程度.相关程度越高,信任度应该越高.问题是,我们如何表达用户的查询意图.通过对提交到 TRECVID'06 的 76 个自动搜索结果的性能分析,我们发现,尽管搜索列表前列的正例样本比例较小,但它们的比例依然比后面的要大.也就是说,排在最前面的初始搜索结果通常更可能含有与查询意图相关的镜头.在本文方案中,我们利用这些排在最前面镜头来表达用户的查询意图,这种思想部分来源于伪相关反馈方法(PRF).因此,评估对一个帧元素的信任程度转化为计算帧元素中镜头和排在最前列镜头的关联程度,这就涉及到集合之间相似度计算的问题.在我们选择度量方式时,有一个因素不容忽视,即排在最前列的镜头不一定都是和查询相关的.这意味着,采用的集合度量方法应能够处理噪声问题.为此,我们采用 Partial Hausdorff Distance[7]来计算最前列镜头集合和各个帧元素之间的相似度.和传统的 Hausdorff Distance 不同的是,Partial Hausdorff Distance 可以自动地选择最好的几个匹配点来度量两个集合的相似度,所以很适合度量含有噪声点的集合.在我们要度量的镜头集合中,最前列镜头集合中的负例就是噪声点.Partial Hausdorff Distance 的定义如下:

$$PHD(A,B)=\max\{phd(A,B),phd(B,A)\} \tag{1}$$

$$phd(A,B)=\underset{a\in A}{K^{th}}\{\min_{b\in B}\{d(a,b)\}\} \tag{2}$$

$$phd(B,A)=\underset{b\in B}{K^{th}}\{\min_{a\in A}\{d(b,a)\}\} \tag{3}$$

其中,A,B 代表两个待测量的集合,$\underset{a\in A}{K^{th}}$ 表示将 A 中每一个点到 B 中所有点的最小距离按升序排序后的第 K 个值;$\underset{b\in B}{K^{th}}$ 表示将 B 中每一个点到 A 中所有点的最小距离按升序排序后的第 K 个值.

事实上,在本方案中,只有 $phd(A,B)$ 被用来测量两个样本集合之间的距离.其中,A 在此表示排在最前列初始镜头集合,我们选择了前 30 个镜头,并称之为伪相关集合;B 在此表示一个帧元素.对于 K 值的选择,经验地设定为 5.也就是说,在前 30 个镜头中,只有 5 个自动选择的镜头被认为是正例镜头,并参与计算,其它镜头被忽略,从而避免了伪相关反馈遇到的问题.

对于每一个识别帧,我们可以根据帧元素与伪相关集合的距离值来评定对它们的信任等级.根据帧元素个数,我们设置相应的等级个数.在我们实验中,我们设定 3 个等级.在证据空间 A,初始列表被聚为三类后,形成一个含四个元素的识别帧 $FOD_A=\{A_1,A_2,A_3,\phi\}$.利用以下基本信任概率分配函数,我们可以为每一个帧分配一个初始信任概率:

$$m(A_i)=\frac{Rank(A_i)}{\sum_{A_k\in FOD_A}Rank(A_k)} \tag{4}$$

在上式中,$Rank(A_i)$代表我们对帧元素 A_i 的信任等级,它从{1,2,3}中选取,数字越大等级越高.信任等级的选取是由式(2)决定.由式(2),可以分别计算出 A_1,A_2,A_3 和伪相关集合的 Partial Hausdorff Distance,距离越小,信任等级越大,分配的数字越大.这样,经排序确认等级后,我们可以将三个等级数字分配给三个帧元素.由基本信任概率分配函数的定义可知:

$$\sum_{A_k\in FOD_A}m(A_k)=1 \tag{5}$$

如果我们进一步假设

$$m(\phi)=0 \tag{6}$$

则我们定义的基本信任概率分配函数完全满足 Dempster-Shafer 理论的要求.

根据 Dempster-Shafer 理论,我们对某个帧元素的信任值是由信任函数决定的,其定义如下:

$$Bel(C)=\sum_{B|B\subseteq C}m(B) \tag{7}$$

即对于某个集合 C 的信任程度是由它所包含的所有子集合的初始信任概率的和而获得的.因为在我们的实际应用中,不同的识别帧都是通过对同一个初始列表的不重叠划分得到的,所以

$$Bel(C)=m(C) \tag{8}$$

到此为止,我们利用信任分配函数为每一个识别帧的每一个元素分配了一个信任值,也就是为多证据融合做好了准备.图 1 显示了这一信任值分配过程.

2.3 融合多个证据

经过以上预处理步骤,对于同一个问题,在每一个证据空间都为其形成一个识别帧,并且为帧中每一个断言分配了一个信任值.我们的目标是取得一个唯一的、质量改善的结果列表.因此,关键的问题是如何有效地融合多方证据,从而将最可能相关的镜头排在最前列.在这一小节,我们将详细介绍如何利用 Dempster-Shafer 证据融合理论,来协同推断最相关镜头.Dempster-Shafer 证据融合规则的核心思想是只保留那些证据一致的集合,而完全忽略证据冲突的集合.Dempster-Shafer 证据融合规则的定义如下所示:

$$m_{1,2}(C) = \frac{\sum_{A\cap B=C}\{m_1(A)m_2(B)\}}{1-K} \quad (9)$$

$$K = \sum_{A\cap B=\phi}\{m_1(A)m_2(B)\} \quad (10)$$

在式(9)、(10)中,K 测量了两方证据对集合 C 的冲突程度.$(1-K)$ 意味着证据冲突的集合完全被排除.在我们的应用中,这意味着 C 中镜头必须是两方都认为相关的镜头.这样就能保证最相关的镜头排在最前面.假设,任意两个来自不同识别帧的元素相交不为空,即 $K=0$.在实际情况下,一些交集可能为空,但为方便描述,我们假设不为空.事实上,这并不影响最后的排序结果,因为无论对空集合的信任值有多高或多低,这个空集合都没有镜头参与排序.注意,$m_1(\cdot)$和$m_2(\cdot)$表示两个不同识别帧的基本信任概率分配函数,而 $m_{1,2}(\cdot)$则是融合后的识别帧的基本信任概率分配函数.

因为不同的识别帧都是通过对同一个初始列表的不重叠划分得到的,所以他们之间的任何两个交集也是不重叠,并且所有交集的集合形成一个新的识别帧.通过式(9),我们就可以为这个新识别帧中的元素分配信任值,并根据信任值来评价帧元素中的镜头和查询的相关性.这样,我们就可以得到一个在集合级的排序.注意,在图 1 中,多证据融合后,新识别帧的所有帧元素的初始信任概率加和不为 1,这是由四舍五入造成的误差.对于那些具有相同信任值的帧元素,我们可以再次利用式(2)来确定它们在列表中排序位置.而在同一个帧元素中的镜头,我们则直接利用它们在初始列表中的等级来排序.其准则如下:

$$rank(S_{i,m}) > rank(S_{i,n}),$$
$$if\ TextScore(S_{i,m}) > TextScore(S_{i,n}), S_{i,m}, S_{i,n} \in U_i \quad (11)$$

其中,$S_{i,m}$ 和 $S_{i,n}$ 代表子集 U_i 中的两个镜头.$TextScore(S_{i,m})$表示镜头 $S_{i,m}$在初始搜索列表的分数或位置,这取决于具体搜索引擎的算法;$rank(S_{i,m})$表示镜头 $S_{i,m}$在 U_i 中的排序,值越大,越靠前.最终,我们得到一个唯一的重新排序的搜索列表.

3 实验结果与分析

为了验证本文建议的视频排序算法的有效性,我们利用 NIST TRECVID'06 视频数据集来测试本算法.此数据集合包括近 343 小时的新闻广播视频,其中有 169 小时的视频作为训练集,其余的 174 小时的视频作为测试集.Fraunhofer Institute[8]则提供了整个视频数据集的镜头分割信息和关键帧.

在我们的实验中,MAP(mean average precision)和 AP(average precision)被用来评价视频搜索的性能.AP 通常用来表示对一个查询话题进行搜索的性能,它不仅能反映出搜索的整体性能,还能很好地反映出正例的分布情况.对于同一个镜头集合,正例镜头靠前的排列的 AP 值要比靠后的排列大的多.MAP 则是对多个查询话题 AP 的平均.

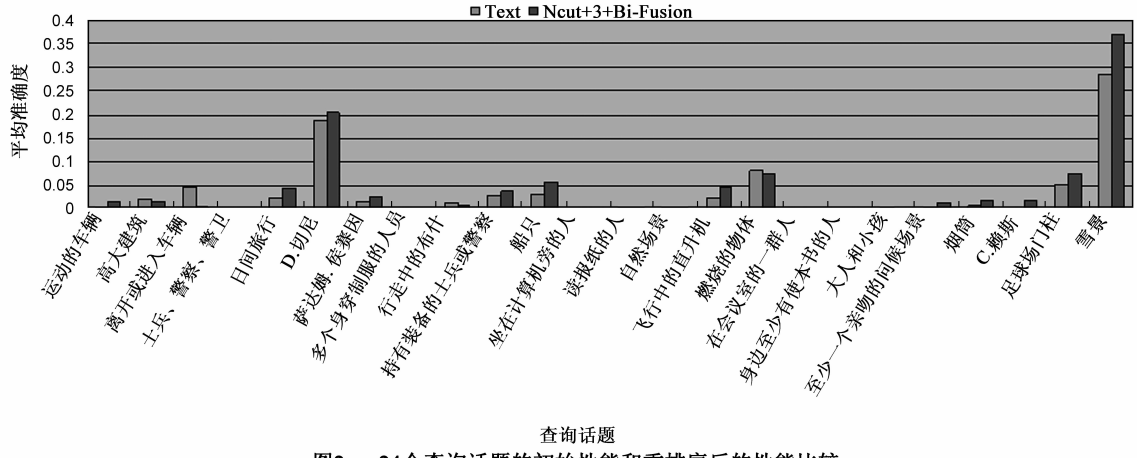

图2 24个查询话题的初始性能和重排序后的性能比较

3.1 自动初始搜索和特征提取

在我们执行排序算法之前,必须先获得一个初始搜索结果列表.为此,我们使用先前开发的基于文字的视频搜索引擎来进行初始搜索.此搜索引擎的基本思想是将视频搜索转化为对视频中语音翻录文字的搜索,其技术细节请参考文献[9].另外,我们还需要提取两个近似独立的特征来构造两个证据空间.因为镜头被视为最小的检索单元,所以特征的提取也是基于视频镜头的.在我们的方案中,我们从镜头的视觉特性和语音文本视角来选择特征.对应镜头视觉特征,我们直接使用文献[10]开发的 120 维的特征矢量;对应文本特征,我们利用我们先前开发的一个 78 维的矢量[9].

3.2 对不同查询话题的性能分析

在这一小节,我们分析了建议的排序算法对不同查询话题的敏感性,并和初始搜索结果进行了对比.在实验中,我们利用 TRECVID'06 使用的 24 个查询话题

进行了测试.图 2 给出了对所有 24 个查询话题执行排序前后 AP 值的统计结果.从图中可以看出,我们的算法对命名的人物和命名的物体效果最好,例如 D. Cheney.事实上,这是一个合理的结果,因为我们采用的文本特征很适合这类话题的查询.另外,对于那些具有显著视觉特征的查询话题,我们的算法也取得了明显的提高,例如雪景.另一方面,对于一些查询话题,例如运动的车辆,重排序的结果反而不如初始搜索的结果.其实,这也是可以理解的,因为使用的特征视角没有运动的信息,所以对于那些含有运动的查询话题,不能取得很好的效果,我们可以引入相应的运动视角来改进.

3.3 和单源证据排序方案的比较

多证据融合排序方法的一个优势就是可以利用不同证据对推断不同查询话题有效的特性,来融合其结果,从而提高排序的性能.为了验证这一方法,我们设计了两个基于单源证据的排序方法,即 NCut + 3 + TextFeature 和 NCut + 3 + VisualFeature,它们分别只使用的文本证据和视觉证据.和建议的多源证据融合排序方法不同的是,单源方法仅仅在一个证据空间下聚类并分配信任值,然后直接在类内重排序镜头,从而得到一个新的列表.表 1 给出了各种重排序算法和初始搜索结果的比较.从表中可以看出,尽管两个单源证据排序方法都取得的不错的性能,分别提高 18% 和 38.4%,但我们建议的多源融合排序算法的性能更优,达到 40.2%.

表 1　各种重排序算法的整体性能比较

不同方法	MAP	提高率
TEXT – ONLY	0.0333	0.0%
NCut + 3 + TextFeature	0.039	18%
NCut + 3 + VisualFeature	0.0461	38.4%
NCut + 3 + Bi-Fusion	0.0467	40.2%

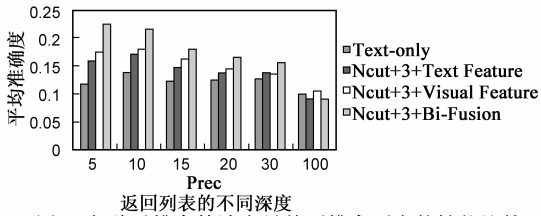

图3　各种重排序算法在最前列排序列表的性能比较

另外,我们建议的重排序算法的另一大优势是,给予排在最前列镜头最高的准确率.图 3 给出了不同排序算法在不同返回深度的搜索性能.注意,这个平均准确度是对所有 24 查询话题的平均.从图中可以看出,在前 100 内的所有深度位置,建议算法的性能都比其它三个优异,特别是前 30 个返回结果,性能提高很明显.这正是我们要达到的目标.

4　结束语

本文提出了一种新颖的、基于多源证据融合的视频排序算法,此算法可以用来重新排列初始搜索结果,将最可能为正例的结果排在最前面.具体来说,初始列表中每一个镜头首先被在多个证据空间独立表示.然后,初始搜索结果列表中的镜头被分别在各个证据空间聚成固定的几类,即形成一个识别帧.利用建议的信任分配函数,每一个帧元素被分配一个信任值.最后,这些来自多源证据的识别帧被我们建议的多源证据协同推断方法融合为一个唯一的结果列表.本算法主要的优点是,不仅整体的性能有所改善,更重要的是给予了排在前列的镜头更高的精度.实验结果也显示了本算法的有效性.

作者简介:

韦世奎　男,1979 年 12 月出生于河北正定.现为北京交通大学信息科学研究所博士研究生,研究方向为模式识别、多媒体分析与检索、拷贝检测等.
E-mail:shkwei@gmail.com

赵　耀　男,1967 年 12 月出生于江苏,教授,博士生导师.现为北京交通大学信息科学研究所所长,北京市"现代信息科学与网络技术"重点实验室主任,2004 年入选教育部新世纪优秀人才支持计划.研究方向为图像编码、数字水印、基于内容的图像与视频检索、多媒体信息处理等.E-mail:yzhao@bjtu.edu.cn

朱振峰　男,1974 年五月出生于黑龙江,博士.2001 年在哈尔滨工业大学获工学硕士学位,2001 – 2005 年在中国科学院自动化研究所模式识别国家重点实验室攻读博士学位.2005 年 4 月进入北京交通大学信息科学研究所工作.研究方向为目标跟踪、多媒体分析与检索、模式识别等.
E-mail:zhfzhu@bjtu.edu.cn

参考文献:

[1] M S Lew, N Sebe, C Djeraba, R Jain. Content-based multimedia information retrieval: state of the art and challenges[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2006, 2(1): 1 – 19.

(下转第 166 页)