一种用于 MAM 的语义可扩展视频编目与检索方法

丁国祥[1),2)] 吴仁炳[2)] 张振亚[1)] 王煦法[1)]

[1)](中国科学技术大学计算机科学与技术系,合肥 230027) [2)](安徽电视台,合肥 230066)

摘 要 针对媒体资产管理系统(media asset management, MAM)对视频检索有着精确定位的特殊要求,提出了一种新的编目与检索方法,该方法结合了基于内容检索和基于手工检索的优点。首先采用主元分析(PCA)的方法对视频图像帧的高维特征进行降维,实现镜头自动分割,并对大量存在的新闻相似镜头进行语义自动标注,然后实现视频编目语义的动态更新与扩展。实验结果表明该方法有效、可行,大大提高了编目工作的效率以及编目语义的质量。

关键词 视频检索 语义 PCA 镜头检测

中图法分类号:TP391 **文献标识码**: A **文章编号**: 1006-8961(2005)08-1036-06

A Method of Semantic Extending to Video Logger and Retrieval for MAM

DING Guo-xiang[1),2)], WU Ren-bing[2)], ZHANG Zhen-ya[1)], WANG Xu-fa[1)]

[1)](*Computer Science & Technology Department, University of Science and Technology of China, Hefei* 230027)

[2)](*Anhui TV Station, Hefei* 230066)

Abstract B To meet the demand of exact localization of video retrieval in Media Asset Management(MAM) system, this paper presents a novel method for logger and retrieval, which takes advantage of both the retrieval based on content and the method based on hand-logger. First of all, the high dimension features are reduced by PCA, which are used to detect the video shots. The semantic of the repeat news shots is auto annotated. The semantic of shots is updated automatically finally. It is verified by the experimental results that the proposed approach is feasible and efficient.

Keywords video retrieval, semantic, PCA, shot detection

1 引 言

媒体资产管理(media asset management, MAM)已经成为广播电视领域近期的热门话题,广播电视机构建立 MAM 系统的主要目的是为了实现数字新闻节目的高效再利用和增值服务,这就要求必须建立一个十分有效的素材索引机制。视音频检索系统是媒体资产管理系统的核心内容之一,同时也是媒体资产管理中最主要的技术难点。

目前媒体资产管理系统中视频检索的研究热点主要体现在以下两个方面:

(1)基于内容的视频检索,即通过建立视频图像底层特征和高层语义之间的联系,实现视频镜头语义的自动标注[1,2],这是视频检索的理想模式之一。但是,电视节目编辑要求精确的检索定位,而视音频底层特征和高层语义之间的鸿沟决定了基于内容的检索在媒体资产管理系统中应用还有一段不小的距离;

(2)以规范化数据为主体的检索方式,利用手工

基金项目:国家自然科学基金项目(60401004);中国博士后科学基金资助项目(2004036463)

收稿日期:2004-11-19;**改回日期**:2005-03-07

第一作者简介:丁国祥(1973~),男。现为中国科学技术大学计算机科学技术系计算机应用专业博士研究生。主要研究方向为智能信息处理、多媒体技术、广播电视技术、模式识别、神经网络等。E-mail:gxding@ustc.edu

对视频镜头进行语义编目,进而实现检索。其主要精力主要集中在编目语法的制订,使得编目语义更具完整性、规范性和一致性。为此,国家有关部门正在加紧制定视音频编目标准。这种检索方式缺点十分明显,编目工作过于繁琐,编目工作的进度往往跟不上新的视频产生的速度,此外,编目语义的时效性和通用性也比较差。

针对以上矛盾,结合以上两个方向研究的优点,提出了一种新的编目与检索系统,主要包含以下 3 个步骤:首先采用主元分析(PCA)的方法对视频序列帧的高维内容特征进行降维,根据降维的特征数据实现镜头的自动分割,分割阈值动态选择,并利用基于子区域的空间结构特征匹配的方法避免了闪光灯事件对镜头分割的干扰;其次,注意到这样一个事实,电视新闻信息存在大量的冗余,如同样的一条新闻在 24 小时内的早间新闻、整点新闻、午间新闻、晚间新闻、新闻联播以及各种新闻专题中滚动播出的同时,还会在同一个电视台的不同频道滚动播出,根据这一现象,利用降维特征数据,实现基于杂凑表的视频语义匹配,并给出了新的视频匹配计算公式,从而实现内容相似镜头语义的自动标注;最后,针对媒体资产管理系统对视频编目与检索的特殊要求,改进文献[1]图像语义网络的结构、初始化输入以及语义动态调节方法,实现视频编目语义的动态更新和扩展,大大提高了编目工作的效率。

2　系统结构

图 1 为用于媒体资产管理系统的视频语义可扩展编目系统的结构图,主要由镜头自动分割、手工编目、编目语义动态扩展与更新 3 部分组成。首先运用 PCA 的方法实现对视频图像帧内容特征进行降维处理,根据降维结果实现镜头的自动分割;然后对新的镜头进行语义匹配,如果能够和语义库内容进行匹配,则对其自动标注相同的语义,否则手工标注;语义动态更新模块根据用户的反馈,实现已编目语义的不断更新和扩展。

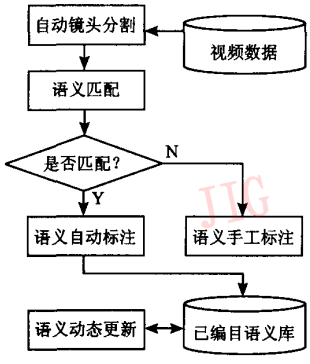

图 1　系统结构图

Fig. 1　The structure of the system

3　镜头自动分割

镜头自动分割[3-5]是视频检索中的难点之一,对新闻视频镜头而言,自动分割的主要难点包括如何有效避免闪光灯干扰以及分割阈值的选择。

3.1　基于 PCA 的降维

视频信息相对于图像来说,是一种时间序列函数,为了综合一段时间上的统计信息,算法中采用主分量分析[6]算法,降低了数据集的维数,同时有效地表达视频序列中基于颜色的动态特征,准确把握住了视频信号的切变点所在,从而实现镜头自动分割。

提取每幅图像帧 Hue 分量的直方图 $\boldsymbol{x}$ 描述图像的颜色特征。$\boldsymbol{x}$ 是一个 d 维矢量,对于一个具有 N 个图像帧的特定视频序列,考察其 H、S、V 分量,可以得到矢量集$\{\boldsymbol{x}_i, i=1,2,\cdots,N\}$,其中,

$$\boldsymbol{x}_i=(h_1,h_2,\cdots,h_d,s_1,s_2,\cdots,s_d,v_1,v_2,\cdots,v_d)$$

其 PCA 算法描述如下:

对原始数据进行均值零化,处理后的数据集对于每一维数来说均值为零:

$$\boldsymbol{x}'=\boldsymbol{x}-\boldsymbol{\mu}$$

式中

$$\boldsymbol{\mu}=\frac{\sum_{i=1}^{N}\boldsymbol{x}_i}{N} \tag{1}$$

计算协方差矩阵 $\boldsymbol{C}$ 及其特征向量 $\boldsymbol{v}_i$ 和特征值 λ_i

$$\boldsymbol{C}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{C}\boldsymbol{v}_i=\lambda_i\boldsymbol{e}_i \tag{3}$$

选择主分量并形成特征矢量集,若原数据集为 d 维,那么可以得出 d 个特征向量 $\boldsymbol{e}_i$ 和 d 个特征值 λ_i。将 λ_i 按降序排列,则前 p 个最大的特征值所对应的特征向量包含了原数据集的大部分信息,因此可以用上述 p 维特征向量集来近似代表原有的数据集。这样,数据集维数可以由 d 维减少为 p 维:

$$\boldsymbol{F}_{\mathrm{Vector}}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p) \tag{4}$$

得出新的数据集

$$\boldsymbol{F}_{\text{Data}} = \boldsymbol{F}_{\text{Vector}}^{\text{T}} \times \boldsymbol{D}_{\text{Adjust}} \tag{5}$$

其中,

$$\boldsymbol{D}_{\text{Adjust}} = (\boldsymbol{x}_i - \boldsymbol{\mu}) \quad 1 \leqslant i \leqslant N \tag{6}$$

得出的新的矢量集为$\{y_1, y_2, \cdots, y_p\}$,对于每幅图像帧来说,其颜色特征可由下式得出:

$$\gamma = \sqrt{y_1^2 + y_2^2 + \cdots + y_p^2} \tag{7}$$

计算相邻图像帧之间的帧差,并且和阈值相比较来判断其是否是镜头的切变点。

3.2 自动分割阈值的选择

阈值选择是镜头分割中的难点问题之一,已经有文献专门对此展开了研究[7,8],并且取得了良好的效果。这里采用了基于局部窗口的方法[7],计算窗口内特征值的均值μ_{window}和方差σ_{window},当方差大于一定值T_2时,选较大的阈值T_{h}(通常为 3 ~ 5 倍均值);反之,当方差比较小时,选较小的阈值T_{s}(通常为 2 ~ 3 倍均值)。

3.3 避免闪光灯干扰

目前已有不少学者提出了避免闪光灯干扰的镜头检测方法[7,9],但它们都是基于闪光灯持续时间很短这样一个假设,然后根据闪光灯位置前后帧的统计特征基本相同的原理来区别闪光灯效果和真正的镜头边界。

随着闪光灯及电池快速充放电等技术的发展,闪光灯的闪光间隔越来越短,长时间持续闪光效果使得帧间图像亮度差值形成连续的峰值,此时,传统的方法就不能准确地识别出闪光灯事件。针对此问题,提出了基于子区域空间结构特征匹配的有效避免闪光灯干扰的镜头自动分割方法,对于上述初步分割的结果,再通过镜头的时间结构特性判断潜在的闪光灯事件位置,最后对潜在位置当前帧和前一帧进行基于子区域的空间结构特征匹配,从而有效消除各种闪光灯效果对镜头检测的干扰。

4 语义自动标注

由于电视新闻的特殊性,存在大量的冗余信息,如果有效地识别这些冗余信息,将会大大提高语义标注的效率。本文利用基于 PCA 的降维信息,对内容相似的视频镜头进行自动识别,并自动标注语义,取得了非常好的效果。

4.1 基于杂凑表的镜头语义自动标注

由于视频数据的海量特性,为了实时实现新输入视频图像帧和视频库中原有的视频图像帧的特征比较,文献[10]、[11]对此类问题提出了基于杂凑表的解决方法。使用该方法成功地对具有相似特征的视频镜头进行了语义自动标注,赋予初始权值后,其结果用作动态语义更新系统的初始化输入。基于杂凑表的插入和字典查询操作可以在常数时间内完成。

考虑到通常情况下,电视新闻在 24 小时后重播的概率非常小,一个电视台一天的新闻总制作量不会超过 10 个小时,每个镜头的平均帧数是 15,则杂凑表长度可以设为

$$l = \frac{10 \times 3600 \times 25}{15} = 60\,000$$

根据上述计算的每帧图像的颜色特征值$\{\gamma_1, \gamma_2, \cdots, \gamma_N\}$,插入图像帧到杂凑表的相应位置中,具有相似特征的镜头帧应该插在同一个链表中。但是由于颜色特征并不能完全表征图像的高级语义,杂凑链表结果中一定含有误匹配镜头,再处理过程将进一步减少这种误匹配,使得语义自动标注结果更加准确。

4.2 视频镜头语义匹配

对于杂凑表中的每一个链表值,进行精确匹配计算,确保同一链表中的镜头确实具有相同的语义。由于每副图像的颜色特征都可以用式(7)来近似描述,对新输入的镜头$\boldsymbol{A}$,语义库中存在的镜头$\boldsymbol{B}$,$\boldsymbol{I}$和$\boldsymbol{X}$分别为镜头$\boldsymbol{A}$、$\boldsymbol{B}$中图像帧,定义

$$E_{I \cap A} = \begin{cases} 1 & \boldsymbol{X} \in \boldsymbol{A} \quad \text{and} \quad \varepsilon > 0, |\gamma_I - \gamma_X| < \varepsilon \\ 0 & \text{其他} \end{cases} \tag{8}$$

$$N_{A \cup B} = \sum E_{X \in A \cup B} = \sum \begin{cases} 1 & \boldsymbol{X} \in \boldsymbol{A} \text{ or } \boldsymbol{X} \in \boldsymbol{B} \\ 0 & \text{其他} \end{cases} \tag{9}$$

视频镜头匹配值S_{sim}如下

$$S_{\text{sim}}(\boldsymbol{A}, \boldsymbol{B}) = \frac{\sum\limits_{I \in (A \cup B)} E_{I \cap A} \cdot E_{I \cap B}}{N_{(A \cup B)}} > \varphi \tag{10}$$

其中,φ为经验设置的匹配阈值。

图 2 给出了匹配实验示意图,圆形点和矩形点分别表示镜头 1 和镜头 2 的序列帧在 2 维特征空间的分布情况,为了更清楚地阐述原理,仅抽样每个镜头的 3 个图像帧,根据式(10)的计算得出其相似度为

$$S_{\text{sim}}(\text{shot}_1, \text{shot}_2) = \frac{1}{5}$$

根据匹配计算结果,按照下式进行语义自动或手动标注,匹配值大于阈值时,自动标注相同的语

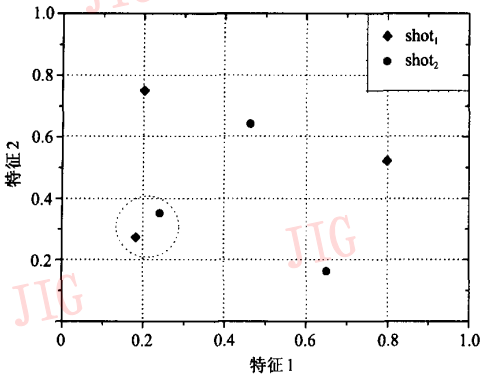

图 2　视频序列帧匹配图

Fig. 2　Diagram of video sequence match

义,否则进行手工标注。

$$\begin{cases} S_B \Rightarrow S_A & \text{如果 } S_{\text{sim}}(\boldsymbol{A},\boldsymbol{B}) > \varphi \\ S_{\text{new}} \Rightarrow S_A & \text{其他} \end{cases} \tag{11}$$

式中,S_A、S_B、S_{new}分别表示镜头 $\boldsymbol{A}$、$\boldsymbol{B}$ 以及手工标注的语义。

5　语义动态更新与扩展

5.1　语义网络

对文献[1]图像语义网络做如下改进。

(1)网络结构　新的视频语义网络结构如图 3 所示,是一个 3 层的结构,分别是关键词层、语义层以及镜头层,具有相同语义的镜头在同一个链表中,语义和视频镜头之间建立动态权值。这样的结构非常适合基于杂凑表的视频镜头匹配。

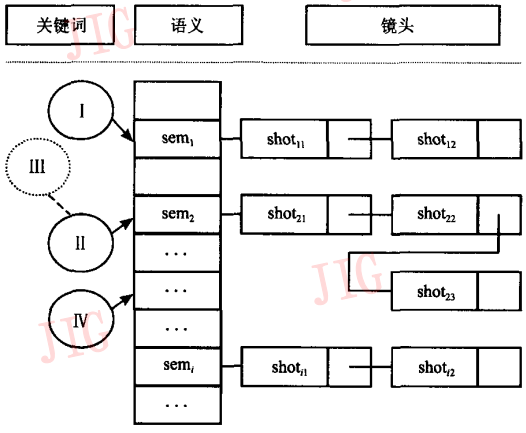

图 3　语义更新与扩展示意图

Fig. 3　Extending and updating of video semantic

(2)初始化输入　在文献[1]中,动态语义更新的初始化输入通常是对少量的样本进行手工标注,然后通过内容特征的计算和比较,对相似的图像进行自动标注。在规模不大的时候,不失为一个比较好的方案,但是在媒体资产管理系统的海量信息中,要想取得比较好的效果,样本数量也可能是海量的,手工初始化是不现实的。本文充分利用电视信息的冗余度非常大的特点,对大量的冗余信息按照式(11)进行初始化赋值。

(3)语义动态调节方法　对于改进的网络结构和初始语义输入方法,提出如下语义动态调节方法:

定义视频镜头语义结构如下:

$$\boldsymbol{S}_i^{\mathrm{T}} = \boldsymbol{w}_i \boldsymbol{T}_i^{\mathrm{T}}$$

其中,$\boldsymbol{w}_i = (w_{i1}, w_{i2}, \cdots, w_{in})^{\mathrm{T}}$ 为语义的权重,$\boldsymbol{T}_i = (T_1, T_2, \cdots, T_n)$是内容描述词向量,$\boldsymbol{S}_i = (s_{i1}, s_{i2}, \cdots, s_{in})$为镜头内容语义。

在图 3 中,Ⅰ、Ⅱ、Ⅲ、Ⅳ是镜头的关键词描述,在初始化阶段,假设关键词Ⅲ没有和相应的镜头对应。当用户输入关键词Ⅲ后,没有用户满意的结果输出;用户再次输入和Ⅲ相关联的关键词(如Ⅱ),并按下式[1,12]转移查询向量,系统返回的结果中一定含有用户满意的结果。

$$\boldsymbol{Q}' = \boldsymbol{Q}_{\text{new}} + \alpha \boldsymbol{Q} + \beta \left(\frac{1}{N'_R} \sum_{i \in S'_R} \boldsymbol{S}_i \right) - \gamma \left(\frac{1}{N'_N} \sum_{i \in S'_N} \boldsymbol{S}_i \right) \tag{12}$$

式中,$\boldsymbol{Q} = \{q_j \mid j = 1, 2, \cdots\}$ 为查询向量,$\boldsymbol{Q}_{\text{new}}$为新输入的查询向量,$\alpha, \beta, \gamma$ 为调节参数,N'_R, N'_N 分别表示输出的查询结果中正反例个数,S'_R, S'_N 分别为输出的正例和反例的语义。

系统在用户感兴趣的镜头和关键词Ⅲ之间建立新的联接,并调整Ⅱ的语义权值为 $\boldsymbol{Q}'$,从而完成编目语义的更新和动态扩展。

5.2　语义更新与扩展算法

算法流程如图 4 所示,主要由初始化模块、查询向量处理模块、交互反馈处理模块、语义调整模块等组成。初始化模块主要是对语义和镜头联接权赋以相同权值;查询向量处理主要包括索引查询词对应的镜头、转移查询向量以及建立多次输入的查询向量之间的关系等;交互反馈是指系统反馈用户查询目标,用户向系统反馈正反例;语义调整是指系统根据用户反馈的正反例增强或降低语义权值,并且输出新的语义结构。

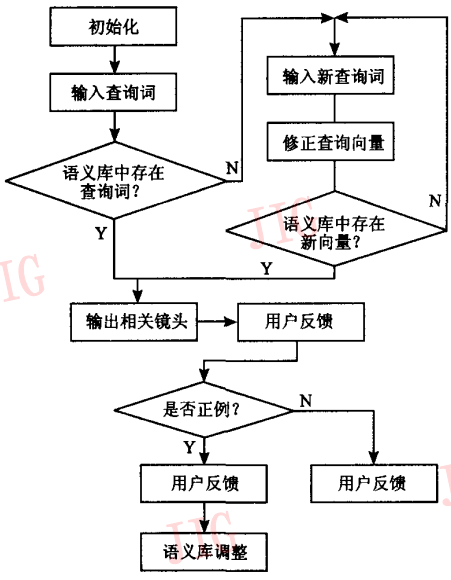

图 4 语义更新与扩展算法流程图

Fig. 4 Flow diagram of semantic updating and extending

算法伪代码如下:

Initialization: $S_{i0}^{T} = W_{i0} T_{i0}^{T}, w_{i0j} = 1, j = 1, 2, \cdots n$

For query: $Q = \{q_j \mid j = 1, 2, \cdots\}$ ($q_j \in \{T_i\}$, or $q_j \notin \{T_j\}$)

While if ($q_j \notin \{T_i\}$)

{ Input new query words: Q_{new}

Shift the query vector as equation(9)

}

Output positive example

Revise the weight as:

$$\begin{cases} w_{ijnew} = w_{ijold} + \delta & \text{if output positive example} \\ w_{ijnew} = \dfrac{w_{ijold}}{\lambda} & \text{if output negative example} \end{cases}$$

Output new semantic:

$S_i = \{w_i, T_i\}$

End

算法中,λ 和 δ 分别为衰减因子和增强因子,目的是对反馈的反例语义权值进行衰减,正例语义权值进行增强;$T_{inew} = (T_1, T_2, \cdots, T_n, q_1, q_2, \cdots)$ 是语义权值调整后新的语义描述。

6 实验系统设计

实验随机选取一天内近 4 个小时的新闻素材,共360 000帧,约4 000个电视节目镜头。

如图版 I 图 1 所示,为两段视频片断的镜头检测结果,每副图中左边是视频播放区,右边是检测出的切变镜头关键帧,同时标出关键帧的位置。

图版 I 图 2 为基于 web 的系统检索界面,用户输入关键词,系统按相似度的大小返回对应视频镜头的关键帧,点击关键帧的图片即播放相应的视频片断。随着用户和系统的不断交互,系统根据交互反馈的结果增加新的语义项,并且各语义项的权重也不断得到调整,从而使得视频数据库中镜头的语义得以动态更新。

图版 I 图 3 是镜头语义自动标注占整个镜头数目的比例,不难看出,对应一个频道的新闻,自动标注镜头的比例接近 30%;而对应不同频道的新闻,其比例接近 60%,这充分说明了电视新闻高冗余度的特性。

图版 I 图 4 是两种算法(本文算法和基于手工编目检索的方法)的平均检索精度-召回率比较,由于手工编目语义通用性差,不同的用户对相同镜头帧的语义理解有差距,显然平均检索结果比本文方法的结果要差。

7 结 论

针对媒体资产管理系统对视频检索的精确定位的特殊要求,提出了一种新的用于媒体资产管理系统的视频编目与检索方法,主要由基于 PCA 的特征降维、镜头自动分割、语义自动标注以及语义自动更新 4 个部分组成。系统结合了基于内容的视频检索和手工编目检索的优点,对大量的冗余信息进行语义自动标注,通过对用户反馈知识的学习,实现了视频语义的动态扩展,既能满足电视节目编辑的精确定位要求,又能最大限度地减少人工的参与,大大提高视频语义编目工作的效率。

参考文献(References)

1 Zhu Xingquan, Zhang Hongjiang. New query refinement and semantics integrated image retrieval system with semiautomatic annotation scheme[J]. Journal of Electronic Imaging, 2001, **10**(4): 850 ~ 860.

2 Bertini M, Bimbo A Del, Pala P. Indexing for reuse of TV news shots [J]. Pattern Recognition, 2002, **35**(3): 581 ~ 591.

3 Uillas Gargi, Rangachar Kasturi, Susan H Stryer. Performance characterization of video shot change detection methods[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2000, **10**(1): 1 ~ 13.

4 Bouthemy P, Gelgon M, Ganansia F. A unified approach to shot change detection and camera motion characterization[J]. IEEE

Transactions on Circuits and Systems for Video Technology, 1999, **9**(7):1030 ~ 1044.

5 Yang Na, Luo Hang-zai, Xue Xiang-yang. A method to detect anchorperson shots for digital TV news[J]. Journal of Software, 2002, **13**(8):1559 ~ 1567. [杨娜, 罗航哉, 薛向阳. 一种用于电视新闻节目的播音员镜头检测算法[J]. 软件学报, 2002, **13**(8):1559 ~ 1567.]

6 Ferré Louis. Selection of components in principal component analysis: A comparison of methods[J]. Computational Statistics & Data Analysis, 1995, **19**(6):669 ~ 682.

7 Zhang Dong, Qi Wei, Zhang Hong-Jiang. A new shot boundary detection algorithm[A]. In: Proceedings of 2nd IEEE Pacific-Rim Conference on Multimedia[C], Beijing, China, 2001:24 ~ 26.

8 Cheng Yong, Yang Xu, Xu De. A method for boundary detection with automatic threshold[A]. In: Proceedings of IEEE Region 10 Technical Conference on Computers, Communication, Control and Power Engineering[C], Beijing, China, 2002:582 ~ 585.

9 Yeo B L, Liu B. Rapid scene analysis on compressed video[J]. IEEE Circuits Systems Video Technol, 1995, **5**(6):533 ~ 543.

10 Sabharwal C L, Bhatia S K. Perfect hash table algorithm for image databases using negative associated values[J]. Pattern Recognition, 1995, **28**(7):1091 ~ 1101.

11 Oostveen J C, Kalker A A C, Haitsma J A. Visual hashing of digital video: applications and techniques[A]. In: Proceedings of the International Society of Optical Engineer, Applications of Digital Image Processing XXIV[C], San Diego, 2001:121 ~ 131.

12 Rui Y, Huang T S. A novel relevance feedback technique in image retrieval[A]. In: Proceeding of the 7th ACM International Conference on Multimedia[C], Orlando, Florida, United States, 1999:67 ~ 70.

图1 镜头检测结果

Fig.1 The result of shot detection

图2 检索界面

Fig.2 Retrieval interface

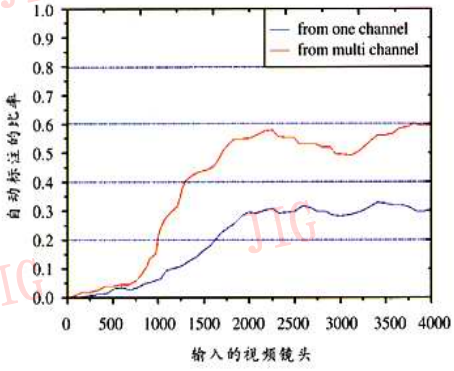

图3 自动标注语义的镜头比例

Fig.3 Ratio of the number of auto-annotation

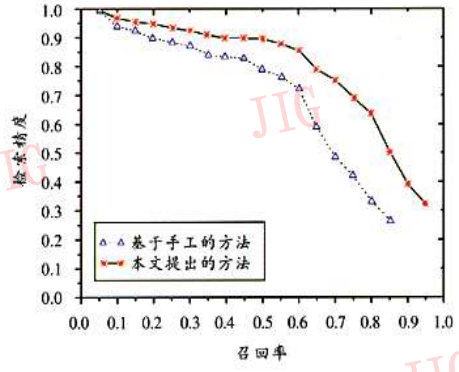

图4 两种算法的平均检索精度-召回率

Fig.4 Average precision-recall on two algorithms