基于语义匹配的交互式视频检索框架

李华北[1] 胡卫明[1] 罗 冠[1]

摘 要 近年来基于内容的视频检索技术受到人们越来越多的关注. 本文提出了一套基于语义匹配的交互式视频检索框架, 其贡献主要为以下三方面: 1) 定义新型的视频高层特征 — 语义直方图用以描述视频的高层语义信息; 2) 使用主导集聚类算法建立基于非监督学习的检索机制, 用以降低在线计算复杂度和提高检索效率; 3) 提出新型的相关反馈机制 — 基于语义的分支反馈, 该机制采用分支反馈结构和分支更新策略实现检索性能的提升. 实验结果表明了本框架的有效性.

关键词 语义匹配直方图, 基于非监督学习的检索机制, 基于语义的分支反馈
中图分类号 TP391.3

A New Interactive Video Retrieval Framework Using Semantic Matching

LI Hua-Bei[1] HU Wei-Ming[1] LUO Guan[1]

Abstract Content-based video retrieval (CBVR) has attracted increasing interest in recent years. In this paper, we propose a new interactive video retrieval framework using semantic matching. The main contributions are three-fold: 1) We define a novel high-level feature named semantic-matching histogram (SMH) to reflect videos′ semantic information. 2) We set up an unsupervised learning-based retrieval mechanism using the dominant set clustering for the sake of low online complexity and high retrieval efficiency. 3) We establish a new interactive mechanism called semantic-based relevance feedback (SBRF) working together with SMHs to improve retrieval performances. Experimental results on a database of sports videos show the effectiveness and efficiency of the proposed framework.

Key words Semantic-matching histogram (SMH), unsupervised learning-based retrieval, semantic-based relevance feedback (SBRF)

基于内容的视频检索 (Content-based video retrieval, CBVR) 已成为多媒体分析领域中最热门的课题之一. 其摆脱了依赖人工文本标注的传统方式, 直接对视频数据所蕴涵的物理和语义内容进行分析和理解, 以期达到快速准确的检索效果. 典型的 CBVR 系统通常包含以下三个关键问题: 视频表征、检索机制和相关反馈.

视频表征是指提取对象特征并定义对象间的相似度. 从结构上讲, 视频不仅是由图像帧组成的简单序列, 更是由关键帧、镜头、组和场景构成的层次结构[1−2]. 从内容上看, 视频的内容主要由如下三种信息表达, 即底层视觉信息、中层序列信息和高层语义信息. 因此, 在哪个信息层次提取视频特征又在哪个结构层次定义相似度便成为视频表征阶段首要解决的问题. Chang 等[3] 使用个体关键帧代表一段视频, 该方法依赖单一图像的视觉信息而完全忽略了视频的序列信息. Dimitiova 等[4] 将两段视频对应帧之间的平均逆距离定义为他们的相似度. 这种方法利用更多的图像试图部分反映序列信息, 然而如何确定视频间图像帧的对应关系成为又一难题. 吴翌等[5] 为每一镜头提取关键帧而后定义镜头质心向量以代表整个镜头, 该方法大大减少了关键帧特征的存储量但是仍然损失了大部分序列信息. Muneesawang 等[6] 使用基于模型的方法提出 “iARM” 系统, 该系统定义了 T-bin 直方图 (T-bin histogram, TBH) 对视频序列信息进行精确建模. 此外, 文献 [7] 进一步挖掘图像高层语义特征以期改进图像检索的性能. 然而, 上述方法的共同问题就是它们对描述视频高层语义信息无能为力.

对于给定查询请求, 检索机制决定着返回相关视频的策略和过程. 传统的检索机制按照视频相似度对数据库进行直接地全排序, 因此又称为直接检索机制. 直接检索机制由于其实施简单灵活得以广泛使用. 然而, 对于每一查询请求, 该机制要求所有相似度计算和比较操作都要在线进行, 势必导致很高的在线计算复杂度和较低的检索效率.

图像检索相关反馈为视频检索提供了基本思想及技术. 然而, 视频检索相关反馈技术发展仍然较为有限. 视频的序列特性决定了用户不得不花费更多的时间来理解输出视频并给出反馈意见. 因此, 视频检索相关反馈技术更加强调交互的效率. 基于底层

收稿日期 2007-06-12 收修改稿日期 2008-04-07
Received June 12, 2007; in revised form April 7, 2008
国家自然科学基金 (60520120099, 60672040, 60705003), 国家高技术研究发展计划 (863 计划) (2006AA01Z453) 资助
Supported by National Natural Science Foundation of China (60520120099, 60672040, 60705003) and National High Technology Research and Deveopment Program of China (863 Program) (2006AA01Z453)
1. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
DOI: 10.3724/SP.J.1004.2008.01243

内容的交互技术[8−9] 似乎并不能满足视频检索的这一要求, 语义相关反馈技术已成为当前研究的热点. He 等[10] 根据用户的反馈信息推导出语义空间, 并通过积累用户意见改进检索性能. 然而该方法过度依赖反馈正例而没有充分利用反馈负例所携带的有用信息. Heisterkamp[11] 将数据库和反馈输出分别看作 “词汇表” 和 “文档”, 采用潜在语义索引捕捉隐藏在查询间的语义信息. 该方法引入新视角来解决相关反馈问题, 但其本质上仍然是一个查询点移动的方法且依赖于大量反馈历史数据. 文献 [12] 使用语义关键词关联的方法建立了基于语义网络的相关反馈. 该方法的主要问题在于如何在缺少适当高层特征的情况下获得初始的语义关联.

针对以上提到的视频检索系统的三个关键问题, 本文提出了一套基于语义匹配的交互式视频检索框架. 首先, 定义新型的高层视频特征 — 语义直方图 (Semantic-matching histogram, SMH) 以描述视频的基本语义内容; 其次, 采用主导集聚类算法, 建立基于非监督学习的检索机制用以降低在线计算复杂度; 最后, 提出新型的交互机制 — 基于语义的分支反馈结构 (Semantic-based relevance feedback, SBRF) 以校正 SMH 所标记的错误主题, 提高检索的性能. 第 1 节详细介绍交互式视频检索框架; 第 2 节给出实验过程及结果; 第 3 节对本文进行总结.

1 交互式视频检索框架

本文致力于通过定义新的视频特征、制定新的检索机制和交互机制来建立一套新型的基于内容的视频检索框架. 如图 1 所示, 该框架由离线操作和在线操作两部分组成. 离线操作包括特征提取和非监督学习两个环节. 首先, 对视频对象提取高层语义特征即语义直方图 (SMH). 随后, 使用 SMH 对整个数据库进行非监督学习. 在线操作包括检索和交互. 检索环节提供直接检索和基于非监督学习的检索两种模式. 最后, 基于语义的分支反馈与 SMH 配合使用, 校正 SMH 标记的不恰当主题, 提升系统的检索性能.

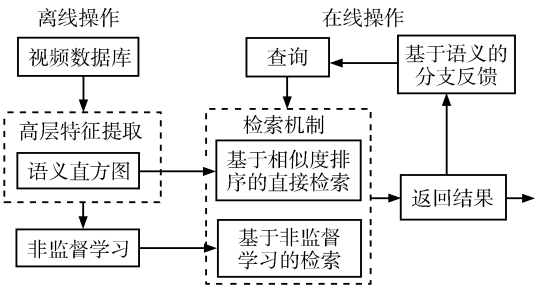


图 1 交互式视频检索框架结构图

Fig. 1 Interactive video retrieval framework

1.1 语义直方图

视频表征是视频检索系统的首要环节. 其中, 高层语义特征能够部分反映视频的语义内容, 更加接近人类的认知习惯. 考虑到同类视频的语义共性, 我们可以简单地使用视频主题对其进行索引, 例如使用 “篮球”、“足球”、“网球” 等体育主题来索引体育视频. 此类主题可以理解为视频的基本语义内容. 具有相似语义内容的两段视频通常包含相似的图像, 视频内容的相似程度往往取决于相似图像的重合程度, 即各段视频引用某一特定典型图像集合的频率. 如果此典型图像集合中的每帧图像都承载着某一预先标定的主题, 那么就可以设法计算出这段视频属于各个主题的概率. 根据这一思路, 我们定义了一种新型的高层语义特征 — 语义直方图用以反映视频的主题. 该特征根据语义匹配策略生成, 具体提取过程分为如下四个阶段: 训练集生成, 模型集生成, 语义匹配和语义直方图提取.

1.1.1 训练集生成

训练集是生成模型集的基础并为整个语义匹配过程提供监督信息, 因此应具备如下条件: 1) 训练样本的主题需要人工标注; 2) 训练样本应为数据库的一个规模相对较小的子集; 3) 在当前底层特征描述能力允许的前提下, 训练集应涵盖尽可能多的主题. 由于底层特征的描述能力有限, 主题对于同种运动的不同场景并不鲁棒. 这就意味着, 如果要使用户能够检索任意体育查询, 那么对应于每种运动的每类场景, 训练集都应该至少包含一个此类样本. 然而, 本节所关注的并不是底层特征, 而是一种学习语义信息的通用方法. 所以, 我们对一种体育主题暂时只选取单一场景的视频来构建数据库, 使得当前底层特征工作良好. 构造语义集 SS 来涵盖数据库中的全部语义主题; 再从数据库中为 SS 中的每一主题选择少量典型视频; 最后提取所有选定视频的每帧图像的底层特征组成训练集 TS

$$\mathrm{SS} = \{\mathrm{Topic}_t\},\ t = 1, \cdots, L \tag{1}$$

$$\mathrm{TS} = \{\boldsymbol{x}_i, \mathrm{Topic\_}x_i\},\ i = 1, \cdots, V \tag{2}$$

其中, Topic_t 表示一个主题, L 表示数据库中的主题总数, $\boldsymbol{x}_i$ 表示承载主题 $\mathrm{Topic\_}x_i \in \mathrm{SS}$ 的图像帧的底层特征向量, V 是训练样本数. 显然, 属于同一段视频的不同帧特征 $\boldsymbol{x}_i$ 具有的主题 $\mathrm{Topic\_}x_i$ 是相同的, 所以 V 远大于 L.

1.1.2 模型集生成

训练集 TS 涵盖了数据库中所有的主题, 每一主题又包含大量的图像帧作为训练样本. 此时我们希望找到一个更加简洁的集合, 在该集合中的每一主题只由少量典型图像来代表. 模型集就是这样的

一个集合, 其中每个模型表示一个从 TS 中提炼出的具有代表性的底层特征向量同时承载某一特定主题. 所以模型集 MS 可视为对训练集甚至整个数据库的精炼, 作为后续语义匹配的模版

$$\mathrm{MS}=\{\boldsymbol{m}_i,\mathrm{Topic\_}m_i\},\ i=1,\cdots,T \tag{3}$$

其中, $\boldsymbol{m}_i$ 表示第 i 个模型, $\mathrm{Topic\_}m_i\in\mathrm{SS}$ 表示该模型的主题, T 为模型总数. 由于某一主题由若干模型表征, 显然 T 也大于 L. 我们改进了竞争学习算法[13], 通过如下步骤生成模型集.

步骤 1. 初始化. 设定模型数远远小于训练样本数, 即 $T\ll V$; 模型初始为零向量.

步骤 2. 选择目标模型. 每次迭代时, 从 TS 中随机选择一个训练样本 $\boldsymbol{x}_k$, 根据下式从当前模型集 MS 中找到与 $\boldsymbol{x}_k$ 最匹配的模型 $\boldsymbol{m}_{i^*}$

$$\|\boldsymbol{x}_k-\boldsymbol{m}_{i^*}\|<\|\boldsymbol{x}_k-\boldsymbol{m}_i\|,\ i=1,\cdots,T,\ i\neq i^* \tag{4}$$

步骤 3. 更新主题. $\boldsymbol{m}_{i^*}$ 和 $\boldsymbol{x}_k$ 之间具有最大的相似度表明它们很可能具有相同的主题. 因此, 将目标模型 $\boldsymbol{m}_{i^*}$ 的主题直接更新为

$$\mathrm{Topic\_}m_i^*=\mathrm{Topic\_}x_k \tag{5}$$

步骤 5. 更新模型. 根据下式更新选定的目标模型 $\boldsymbol{m}_{i^*}$

$$\boldsymbol{m}_{i^*}(j+1)=\boldsymbol{m}_{i^*}(j)+l(j)(\boldsymbol{x}_k-\boldsymbol{m}_{i^*}(j)) \tag{6}$$

其中, j 表示当前迭代次数; J 和 $l(0)$ 分别表示最大迭代次数和初始学习步长; $l(j)$ 为当前的学习步长, 按下式单调下降

$$l(j+1)=l(0)\left(1-\frac{j}{J}\right) \tag{7}$$

1.1.3　语义匹配

语义匹配是将给定视频的每帧图像映射到模型集上, 从而得到最佳匹配主题序列. 如图 2 ($N=3$) 所示, 匹配过程分为三个步骤. 首先, 为给定视频的每帧图像提取底层特征, 将视频帧序列转化为底层特征序列; 然后, 从模型集中为每帧图像找出与其底层特征向量最匹配的 N 个特定模型, 这样又将一条底层特征序列转化为 N 条最佳匹配模型序列; 最后, 根据每个模型的主题将 N 条最佳匹配模型序列映射为 N 条最佳匹配主题序列, 从中即可提取语义直方图.

1.1.4　语义直方图

从 N 条最佳匹配主题序列中, 定义语义直方图 (SMH) 的特征向量如下

$$\boldsymbol{SMH}=[(\mathrm{Topic}_t,p_t)],\quad t=1,\cdots,L \tag{8}$$

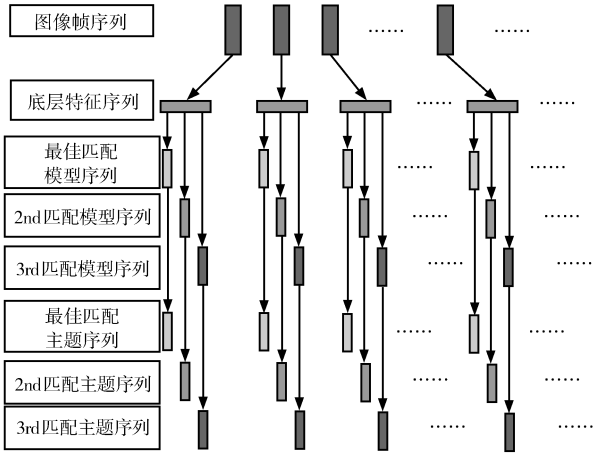


图 2　语义匹配过程, $N=3$

Fig. 2　Process of semantic matching, when $N=3$

其中, Topic_t 表示语义集 SS 中的一个主题, p_t 表示该视频属于主题 Topic_t 的概率. $\boldsymbol{SMH}$ 具有如下优点: 1) 较低的特征维度. $\boldsymbol{SMH}$ 的维度等于主题数 L, 通常不大; 2) 明确的物理意义. $\boldsymbol{SMH}$ 的每一维表示该视频属于相应主题的概率, 因此对应于 $\boldsymbol{SMH}$ 峰值的主题即可作为该视频的最相关主题标记; 3) 稀疏性. 一段视频通常只与有限数目的主题有关, 因此 $\boldsymbol{SMH}$ 通常是稀疏的. 其稀疏性节约了存储空间同时降低了匹配的计算复杂度.

1.2　基于非监督学习的检索机制

本节中, 我们采用主导集聚类算法[14] 建立基于非监督学习的检索机制. 通过离线聚类, 该机制能够达到较低的在线计算复杂度, 同时实现对数据库更为有效的管理. 主导集聚类算法是一种新型的基于图理论的聚类方法, 可以获得较高的聚类质量并且自动确定聚类个数. 主导集聚类的实现主要分为以下两个阶段: 邻接矩阵表示和二次优化问题求解.

1.2.1　邻接矩阵表示

将视频数据库看作一个无向边权图 $G=(V,E)$, V 中的一个节点表示一段视频, E 中的一条边连接一组视频对. 将一组视频对间的相似度定义为这条边的权值. 计算邻接矩阵 $A=[Sim_{ij}]$ 来代表整个边权图 G, 其中 Sim_{ij} 表示视频对的相似度. 邻接矩阵计算是一个相当耗时的过程, 正是该过程的离线执行有效地减少了在线计算的复杂度.

1.2.2　二次优化问题求解

主导集聚类算法等价于如下二次问题

$$\max f(\boldsymbol{u})=\boldsymbol{u}^{\mathrm{T}}A\boldsymbol{u} \tag{9}$$

$$\text{s.t. }\ \boldsymbol{u}\in\Delta=\left\{\boldsymbol{u}\in\mathbf{R}^N|u_i\geq 0\ \text{ and }\ \sum_{i=1}^{N}u_i=1\right\}$$

式 (9) 的局部最优解 $\boldsymbol{u}^*$ 可由如下方程得到

$$u_i(p+1) = u_i(p)\frac{(A\boldsymbol{u}(p))_i}{\boldsymbol{u}(p)^{\mathrm{T}}A\boldsymbol{u}(p)} \tag{10}$$

其中, p 是当前迭代次数. 主导集定义为局部最优解 $\boldsymbol{u}^*$ 中正分量的标号集: 将属于已有主导集的节点从当前图中删除得到新图 G', 重新生成邻接矩阵的子矩阵 A'; 重复求解二次优化问题直至图为空.

对整个数据库进行离线聚类后, 基于非监督学习的查询策略包括以下步骤: 计算每个聚类的聚类中心及方差; 将方差较小的聚类标记为可靠聚类, 将方差较大的聚类重新拆分为自由样本; 计算查询向量与各可靠聚类中心的相似度, 并找到若干相关聚类; 将相关聚类和自由样本组成目标子库; 重新计算查询向量和目标子库中样本的相似度; 对目标子库排序并返回结果. 该策略大大减少了相似度计算和比较的次数, 从而达到降低在线计算复杂度的目的.

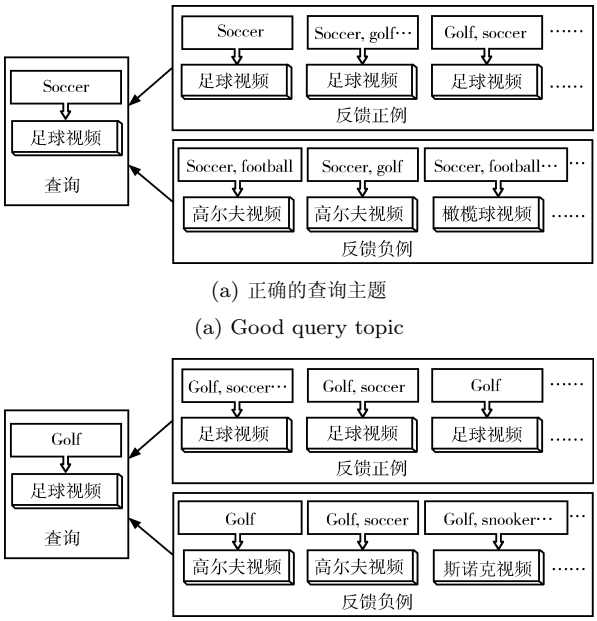
1.3 基于语义的分支反馈机制

传统的相关反馈技术[8−9] 往往独立于具体的特征提取, 具有单一的反馈结构, 并且较少考虑对象的高层语义内容. 针对上述问题, 本节建立了新型的交互机制 — 基于语义的分支反馈 (Semantic-based relevance feedback, SBRF). 该机制与语义直方图 (SMH) 配合使用, 具有独特的分支结构, 通过在线补偿监督信息来校正 SMH 所标记的不恰当的语义主题, 进而提升系统的检索性能.

1.3.1 分支反馈结构

使用语义直方图进行检索时, 查询向量 $\boldsymbol{SMH_Q}$ 能否给出正确的主题标记至关重要. 错误的查询向量主题, 可能导致很差的检索结果. 因此, 本节设计了分支反馈结构用于处理不同的查询主题情况. 图 3 对用户的反馈信息进行分析, 图中每个立方体代表一段视频, 其中的文字表示该视频的真实主题, 视频上方带箭头的方框表示其 $\boldsymbol{SMH}$ 给出的主题列表. 显然, 检索新查询请求时, 系统只能依靠方框中 $\boldsymbol{SMH}$ 所提供的信息, 而并不知道视频的真实主题. 因此在交互过程中, 用户在标记反馈正负样本的同时, 还应该告知检索系统查询向量的初始主题和其真实主题是否一致. 在图 3 (a) 中, 一段 “足球” 查询具有正确的主题 “Soccer”. 所以用户标记的反馈正例必定是一组带有 “Soccer” 主题标记的 “足球” 视频, 而被错误标记为 “Soccer” 的 “非足球” 视频成为反馈负例. 图 3 (b) 给出了 “足球” 查询被错误标记为 “Golf” 的情况. 此时的正例就是同样被错误标记为 “Golf” 的 “足球” 视频, 而负例为具有 “Golf” 主题的 “非足球” 视频.

图 4 为分支反馈结构, 其中查询向量是否具有正确的主题决定了迭代更新的过程. 更新过程包括修改操作和加强操作两部分. 正确的查询主题, 意味着反馈正例的 $\boldsymbol{SMH}$ 向量较为可靠, 而负例的 $\boldsymbol{SMH}$ 不可靠, 例如图 3 (a) 中的情况. 此时, 我们就应该修改不可靠的负例向量同时加强可靠的查询向量和正例向量. 相反情况时, 修改查询向量和正例向量同时适当加强负例向量.



(a) 正确的查询主题

(a) Good query topic

(b) 错误的查询主题

(b) Bad query topic

图 3 对于用户反馈信息的分析

Fig. 3 Analyses of users′ feedback for a “Soccer” query

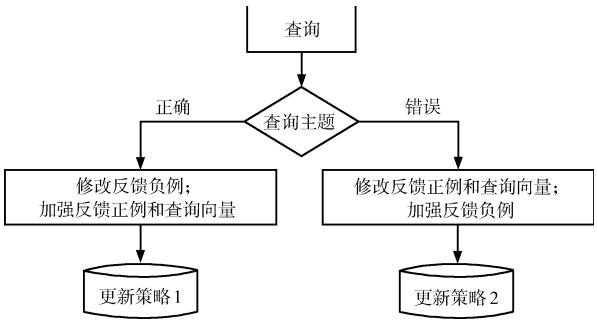


图 4 分支反馈结构

Fig. 4 Branched structure for relevance feedback

1.3.2 分支更新策略

根据上述更新原则, 本节给出具体的分支更新策略如算法 1 所示. 其中 $\boldsymbol{V_Q}$、$\boldsymbol{V_P}$ 和 $\boldsymbol{V_N}$ 分别表示查询请求、反馈正例和反馈负例的 $\boldsymbol{SMH}$ 向量. 设 $Size$ 为向量的维度, $Inc1 \sim Inc3$ 为增量常数. 由于图 3 (a) 中查询向量和正例向量的峰值主题 (Soccer) 得到了用户充分肯定, 所以可以将 $Inc1$ 和 $Inc2$ 设为较大值, 使得算法 1 (a) 中的 $\boldsymbol{V_Q}[Peak\_Position]$

和 $\boldsymbol{V_P}[Peak\_Position]$ 在加强操作和归一化后趋近于 1; 相比之下, 图 3 (b) 中负例向量的峰值主题 (Golf) 仅得到用户的间接肯定, 因此设置相对小的 $Inc3$ 使得算法 1 (b) 中的 $\boldsymbol{V_N}[Peak\_Position]$ 缓慢增大.

算法 1　分支更新策略 (Branched updating strategies)

(a) 对于正确查询主题的更新策略 1 (Strategy 1 for a good query topic)

1) 定位: 找到查询向量 $\boldsymbol{V_Q}$ 的主题位置 $Peak\_Position$, 使得对于任意 $i \neq Peak\_Position$ 有 $\boldsymbol{V_Q}[Peak\_Position] \geq \boldsymbol{V_Q}[i]$.
2) 修改: $\boldsymbol{V_N}$. $\boldsymbol{V_N}[Peak\_Position] = 0$;
3) 加强: $\boldsymbol{V_Q}$,$\boldsymbol{V_P}$.
   a) $\boldsymbol{V_Q}[Peak\_Position] = \boldsymbol{V_Q}[Peak\_Position] + Inc1$;
   b) $\boldsymbol{V_P}[Peak\_Position] = \boldsymbol{V_P}[Peak\_Position] + Inc2$.
4) 归一化并保存更新后的 $\boldsymbol{V_Q}$,$\boldsymbol{V_P}$ 和 $\boldsymbol{V_N}$.

(b) 对于错误查询主题的更新策略 2 (Strategy 2 for a bad query topic)

1) 定位: 同上.
2) 修改: $\boldsymbol{V_Q}$, $\boldsymbol{V_P}$.
   a) $\boldsymbol{V_Q} = \boldsymbol{V_Q} + \sum \boldsymbol{V_P}$
   b) $\boldsymbol{V_Q}[Peak\_Position] = \boldsymbol{V_P}[Peak\_Position] = 0$;
3) 加强: $\boldsymbol{V_N}$.
   $\boldsymbol{V_N}[Peak\_Position] = \boldsymbol{V_N}[Peak\_Position] + Inc3$.
4) 归一化并保存.

图 4 的分支反馈结构和算法 1 的分支更新策略构成了基于语义的分支反馈机制 (SBRF), 其具备如下特点: 1) 具有独特的分支结构用于处理不同的查询情况; 2) 充分利用了反馈正例和反馈负例, 并同时更新反馈样本和查询向量; 3) 与视频高层特征 $\boldsymbol{SMH}$ 相互配合, 从 $\boldsymbol{SMH}$ 中获得视频初始主题并对错误主题进行更新; 4) 随着越来越多的用户参与交互, SBRF 能够不断积累交互所带来的性能提升而无需记录完整的反馈历史数据; 5) 经过足够多次交互, 算法最终收敛于正确的查询主题.

2　实验结果及分析

本节通过实验程序 — "CBVR_System" 来实现上述检索框架. 实验时我们构建了自己的体育视频数据库. 该数据库目前包括 1 024 段视频对象, 分为 32 类体育运动主题, 每类主题包含 32 段视频. 每段视频长度大约 10 秒钟, 可视为一个镜头. 整个数据库总共包含大约 200 000 帧图像, 在这个数据库上我们成功进行了如下实验内容. 实验数据均在 Pentium IV 2.80 G 的微机上测量.

2.1　使用语义直方图的检索效果

底层特征是语义直方图提取的基础, 实验中我们暂时选用颜色相关图[15], 后续工作会加入更多图像描述子如 (纹理、形状等) 以丰富底层帧特征的描述能力. 另外, 提取视频序列特征 T-bin 直方图 (TBH)[6] 用于对比实验. 表 1 给出了各类特征的维度和提取时间. 语义直方图的维度为数据库中主题数, 远小于 T-bin 直方图的维度. 特征提取为离线操作, 所以提取时间并不影响在线检索的实时性. 表 2 给出了使用 TBH 和 SMH 检索的平均性能对比. 查全率、查准率根据前 16 段返回结果进行计算 (此时最高查全率为 50 %). 结果数据为处理 50 段查询的平均值. 使用 TBH 检索的查准率仅为 66.83 %, 平均耗时 1.237 s; 相比之下, SMH 将初始查准率提高到了 86.35 %, 而平均检索时间仅为 0.1 s. 实验结果表明语义直方图具有良好的检索性能.

表 1　特征提取参数

Table 1　Parameters of feature extraction

层次	底层特征	序列特征	语义特征
特征	颜色相关图	T-bin 直方图	语义直方图
维数	192	400	32
提取时间	约 10 h	64 s	93 s

表 2　使用 TBH 和 SMH 检索的平均性能比较

Table 2　Comparison of average precision for retrieval using SMHs and TBHs

	平均查准率 (%)	平均查全率 (%)	检索时间 (s)
使用 TBH 检索	66.83	33.42	1.237
使用 SMH 检索	86.35	43.18	0.100

2.2　非监督学习检索机制的性能

采用主导集聚类算法进行非监督学习时, 需要注意以下两个实际阈值问题, 迭代终止阈值 (IeT) 和分离阈值 (SoT).

2.2.1　迭代终止阈值

求解式 (9) 的二次优化问题时, 式 (10) 迭代终止当且仅当 $u_i(p+1) = u_i(p)$ 对每个 i 成立. 实际上, 满足此条件是相当耗时甚至是不可能的. 实验中, 我们设定迭代终止阈值 IeT, 定义了如下基于阈值的终止条件: $error(p+1) = \sum |u_i(p+1) - u_i(p)|$ 表示 p+1 次迭代后的误差和; $Error\_error(p+1) = |error(p+1) - error(p)|$ 表示两次迭代间误差和的减少量. 当误差和减少量小于 IeT 时, 则迭代终止.

2.2.2　分离阈值

主导集定义为局部最优解正分量的标号集, 即 $\mathrm{DS}_{u^*} = \{i : u_i^* > 0\}$. 然而实验中发现, 局部最优解 $\boldsymbol{u}^*$ 中存在三类分量: 零分量, 不属于本聚类; 较大分量, 构成真实的聚类; 极小分量, 通常为被误分类的类外无关点. 因此, 我们引入分离阈值代替上式中的 "0", 用来滤除极小的类外无关点. 这样, 主导集的定义修改为: $\mathrm{DS}'_{u^*} = \{i : u_i^* > SoT\}$.

确定适当的阈值之后, 使用 TBH 和 SMH 对整个数据库进行非监督学习. 表 3 给出了聚类结果. 其中, 迭代终止阈值 IeT 决定聚类质量和聚类时间, 而聚类数目主要受分离阈值 SoT 影响. 随着 IeT 的减小聚类时间延长; 随着 SoT 的增大聚类数目增多.

表 3 聚类结果

Table 3 Clustering results

	IeT	SoT	聚类时间	聚类数目	编号
使用 TBH 的聚类结果	10^{-11}	10^{-60}	3 min 24 s	19	1
	10^{-11}	10^{-20}	12 min 16 s	57	2
	10^{-14}	10^{-60}	3 min 35 s	11	3
	10^{-14}	10^{-20}	14 min 24 s	111	4
使用 SMH 的聚类结果	10^{-8}	10^{-20}	61s	23	5
	10^{-8}	10^{-10}	2 min 11 s	72	6
	10^{-11}	10^{-20}	8 min 7 s	34	7
	10^{-11}	10^{-10}	24 min 5 s	132	8

表 4 为直接检索和非监督学习检索性能比较 (非监督学习检索选用 4 号和 8 号聚类结果). 相比于直接检索, 基于非监督学习的检索在不影响检索效果的情况下, 可将检索时间缩短到原来的 40 % 左右, 大大提高了检索效率.

表 4 直接检索和非监督学习检索性能比较

Table 4 Comparison between direct retrieval and unsupervised-learning-based retrieval

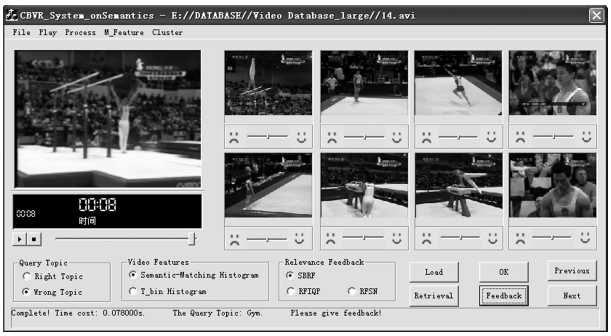
检索机制	特征	查准率 (%)	查全率 (%)	检索时间 (s)
直接检索	使用 TBH	66.83	33.42	1.237
	使用 SMH	86.35	43.18	0.100
非监督学习检索	使用 TBH	66.52	33.26	0.542
	使用 SMH	86.80	43.40	0.039

2.3 基于语义的分支反馈机制的效果

每次交互中, 用户需要完成两项操作: 1) 通过选择界面中的 “Right Topic” 或 “Wrong Topic” 来评估查询向量的主题; 2) 通过拖动视频下方的滑动条来标记反馈样本. 图 5 给出了 SBRF 对于一个错误查询主题的反馈效果. 在图 5 (a) 中, 一段 “体操” 查询被错误的标记为 “American Football”, 导致了极差的检索效果. 即使前两段视频为相关结果, 也是因为它们具有同样的错误标记. 通过 SBRF 后如图 5 (b) 所示, 查询主题得以校正, 检索结果得到显著提高. 表 5 给出反馈前后的性能比较, 其中当返回 16 段结果时 SBRF 将查准率提高到 95.83 %; 当返回 48 段时, 将查全率提高到 91.25 %. 图 6 为使用 SMH 检索经过 SBRF 前后的查全率 — 查准率曲线. 图中每条曲线包含 12 个坐标点, 分别表示返回视频数量从 4 增大到 48 时的查全率和查准率. 如表 5 和图 6 所示, SBRF 通过较少次交互就达到了显著的性能提升.

(a) 反馈前检索结果

(a) Initial retrieval result



(b) 反馈后检索结果

(b) Result after SBRF

图 5 SBRF 对于错误查询主题的反馈效果

Fig. 5 Effect of SBRF for a wrong query topic

表 5 反馈前后的性能比较

Table 5 Comparison before and after SBRF

结果数	反馈前 (%)		反馈后 (%)	
	查准率	查全率	查准率	查全率
4	92.50	11.56	100	12.50
16	86.35	43.18	95.83	47.92
32	79.06	79.06	84.69	84.69
48	56.67	85.01	60.83	91.25

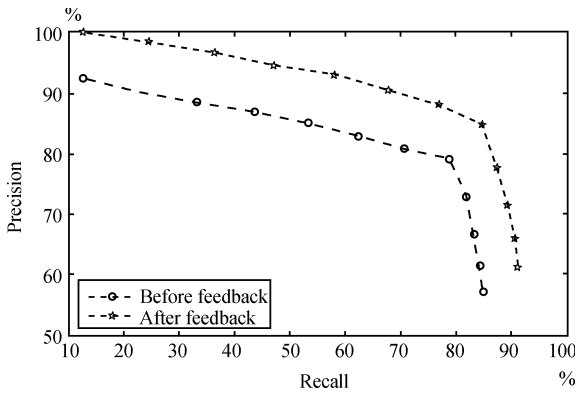


图 6 使用 SMH 检索经过 SBRF 前后的 P-R 曲线

Fig. 6 Precision-recall graph for retrieval using SMHs before and after SBRF

3 结论

本文提出基于语义匹配的交互式视频检索框架, 主要贡献在于将新的视频特征、检索机制和反馈技术集成为一套新型的基于内容的视频检索框架. 其中, 语义直方图依据监督信息通过语义匹配策略, 标记视频可能所属的主题, 挖掘视频的基本语义内容; 基于非监督学习的检索机制可获得较低的在线计算复杂度和较高的检索效率; 基于语义的分支反馈技术采用分支反馈结构和分支更新策略来校正不恰当的视频主题, 提高检索性能. 以上结论均已得到实验验证, 并为我们今后进一步工作铺平了道路.

References

1 Zhang S H, Zhang Y F, Chen T, Hall P M, Ralph Martin. Video structure analysis. *Tsinghua Science and Technology*, 2007, **12**(6): 714−718

2 Shi L, King I, Lyu M R. Video summarization by video structure analysis and graph optimization. In: Proceedings of IEEE International Conference on Multimedia and Exposition. Taipei, Taiwan: IEEE, 2004. 1959−1962

3 Chang H S, Sull S, Lee S U. Efficient video indexing scheme for content based retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 1999, **9**(8): 1269−1279

4 Dimitrova N, Abdel M M. Content-based video retrieval by example video clip. In: Proceedings of International Society for Optical Engineering. New York, USA: SPIE, 1997. 59−70

5 Wu Yi, Zhuang Yue-Ting, Pan Yun-He. Video similarity measurement. *Journal of Computer-Aided Design and Computer Graphics*, 2001, **13**(3): 284−288
(吴翌, 庄越挺, 潘云鹤. 计算机辅助设计与图形学学报, 2001, **13**(3): 284−288)

6 Muneesawang P, Guan L. iARM: an interactive video retrieval system. In: Proceedings of IEEE International Conference on Multimedia and Exposition. Taipei, Taiwan: IEEE, 2004. 285−288

7 Eidenberger H, Breiteneder C. Semantic feature layers in content-based image retrieval: implementation of human world features. In: Proceedings of the 7th International Conference on Control, Automation, Robotics and Vision. Washington D. C., USA: IEEE, 2002. 174−179

8 Kim D H, Chung C W. Qcluster: relevance feedback using adaptive clustering for content-based image retrieval. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. San Diego, USA: ACM, 2003. 599−610

9 Rui Y, Huang T S. Optimizing learning in image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Hitton Head Island, USA: IEEE, 2000. 236−243

10 He X F, King O, Ma W Y, Li M J, Zhang H J. Learning a semantic space from user′s relevance feedback for image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2003, **13**(1): 39−48

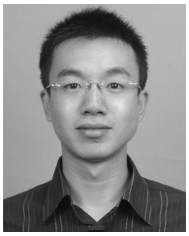
11 Heisterkamp D R. Building a latent semantic index of an image database from patterns of relevance feedback. In: Proceedings of the 16th International Conference on Pattern Recognition. Washington D. C., USA: IEEE, 2002. 134−137

12 Lu Y, Hu C H, Zhu X Q, Zhang H J, Yang Q. A unified framework for semantics and feature based relevance feedback in image retrieval system. In: Proceedings of the 8th ACM International Conference on Multimedia. California, USA: ACM, 2000. 31−37

13 Kohonen T. *Self-Organizing Maps*. Berlin: Springer-Verlag, 1997

14 Pavan M, Pelillo M. A new graph-theoretic approach to clustering and segmentation. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE, 2003. 145−152

15 Huang J, Kumar S R, Mitra M, Zhu W J, Zabih R. Image indexing using color correlograms. In: Proceedings of IEEE International Conference on Computer Vision and Patten Recognition. Washington D. C., USA: IEEE, 1997. 245−268



李华北 中国科学院自动化研究所硕士研究生. 2005 年获得浙江大学控制系自动化专业学士学位. 主要研究方向为多媒体信息检索技术. 本文通信作者.
E-mail: hbli@nlpr.ia.ac.cn
(**LI Hua-Bei** Master student at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He received his bachelor degree from Zhejiang University in 2005. His research interest covers multimedia information processing and retrieval. Corresponding author of this paper.)

胡卫明 博士. 中国科学院自动化研究所模式识别国家重点实验室研究员. 主要研究方向为视频信息处理与网络信息安全识别. E-mail: wmhu@nlpr.ia.ac.cn
(**HU Wei-Ming** Professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interest covers video information processing and recognition of network information security.)

罗 冠 博士, 中国科学院自动化研究所模式识别国家重点实验室助理研究员. 分别于 1998 年、2001 年及 2004 年在西北工业大学电子与信息学院获得工学学士、硕士及博士学位. 2004 年 6 月赴香港城市大学创意媒体学院从事博士后研究工作. 主要研究方向为互联网内容安全视频数据挖掘模式识别和机器学习.
E-mail: gluo@nlpr.ia.ac.cn
(**LUO Guan** Assistant professor at National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He received his bachelor, master, and Ph. D. degrees from Northwestern Polytechnical University in 1998, 2001, and 2004, respectively. He worked as a senior research associate in City University of Hong Kong from June 2004 to August 2005. His research interest covers web content analysis, video data mining, pattern recognition, and machine learning.)