基于语义概念的视频检索系统的设计与实现

刘安文 支 琤 张 瑞 盛骁杰 杨小康
(上海交通大学电子工程系图像通信与信息处理研究所,上海 200240)
(上海交通大学上海市数字媒体处理与传输重点实验室,上海 200240)

摘 要 设计并实现了一种基于语义概念的视频检索系统,该系统包括视频镜头分割与关键帧提取、语义概念检测和用户检索 3 个部分。系统采用镜头分割与关键帧提取对视频进行层次分割,并对关键帧图像提取有效的图像低层特征,再使用支持向量机(SVM)进行概念的检测,最后针对概念内容进行视频检索。在概念检测中,提出了一种基于验证平均准确率的线性加权方法对 SVM 的分类结果进行后融合。实验结果表明,该方法可以达到较高的检索准确率。

关键词 视频检索 支持向量机 概念检测 融合
中图法分类号:TP301.6 **文献标识码**:A **文章编号**:1006-8961(2008)10-2055-04

Design and Implementation of Semantic Concept Based Video Retrieval System

LIU An-wen, ZHI Cheng, ZHANG Rui, SHENG Xiao-jie, YANG Xiao-kang
(*Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai* 200240)
(*Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai Jiaotong University, Shanghai* 200240)

Abstract This paper introduces the design and implementation of a semantic concept based video retrieval system, which consists of shot boundary detection and key frame extraction subsystem, semantic concept detection subsystem and user retrieval subsystem. First, digital video is divided into hierarchical structure for retrieval. Then, efficient low level feature of key frames are extracted. Support Vector Machine is used to detect concepts in such key frames, and the video retrieval is based on those concepts. In the procedure of concept detection, we take a linearly weighted fusion method based on validation precision to improve the average precision. Experiments show that the Mean Average Precision of our system is as high as the best one of all submissions.

Keywords video retrieval, support vector machine (SVM), concept detection, fusion

1 引 言

随着多媒体技术的迅猛发展,数字视频信息量飞速增长。如何有效地在海量视频数据中查找用户感兴趣的内容,已经成为当今信息社会亟待解决的技术问题。开发基于内容的视频检索系统也已成为当今多媒体信息处理领域中一个非常有应用价值和前景的课题。

目前,网络上普遍应用的视频搜索引擎(如 YouTube.com 等)都是沿用传统的文本检索方式,数字视频的检索信息来自于其文本信息,如视频标题、环绕文字和文字简介等。如何使计算机自动从数字视频序列中挖掘出有效的视觉信息并用于检索,是

基金项目:国家自然科学基金项目(60502034,60625103);国家 863 计划项目(1006AA01Z124)
收稿日期:2008-07-11;**改回日期**:2008-07-22

第一作者简介:刘安文(1984~),男。现为上海交通大学电子工程系图像通信与信息处理研究所硕士研究生。主要研究方向为视频检索系统和基于内容的图像分类。E-mail:liuanwen1018@163.com

当前视频检索领域中的一个极具挑战性的工作。

为了鼓励在视频检索领域的研究,从 2003 年起,美国国家标准局每年都在全世界范围内举办视频检索领域权威的学术性比赛 TRECVID,吸引了包括美国哥伦比亚大学、IBM 公司等数十个科研机构参加。语义概念检测(即图像的高层特征提取)正是 TRECVID 的 4 个子任务之一。

本文设计并实现了一种基于视频语义内容的检索系统。该系统不同于传统的视频搜索引擎,用户在检索中是以语义概念(如汽车、人和飞机等)为检索条件,检索结果为视觉内容上包含了相应语义概念的视频镜头。本文在实验中采用了 TRECVID2005 定义的 37 个语义概念作为检索条件,并在 TRECVID2005 发布的视频数据上进行了系统性能的测试。

2　视频检索系统的总体架构

视频检索系统主要由 3 个模块组成,即视频镜头分割与关键帧提取模块、语义概念提取模块和用户检索模块。其总体架构如图 1 所示。

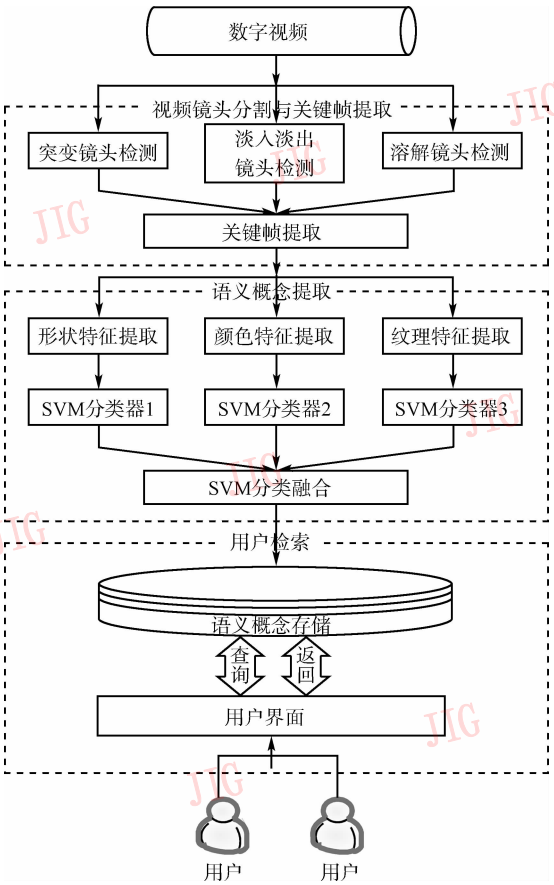

图 1　系统总体架构图

Fig. 1　Architecture of the system

数字视频在检索中应以镜头为检索的基本单位,所以,系统首先在时域上将视频分割为镜头。由于同一镜头由相关性比较强的图像序列组成,所以可以用一幅或者几幅图像帧来表示该镜头,这些图像帧被称作该镜头的关键帧。通过提取关键帧可以去除镜头中冗余的视觉信息。

对得到的关键帧图像,可以提取其颜色、纹理和形状三类图像低层特征,然后对不同的图像低层特征分别构建最优化参数的 SVM 分类器,以进行高层语义概念的检测。对于不同的低层特征的 SVM 分类置信度结果,采用基于验证准确率的线性加权方法进行融合。

最后,将语义概念置信度值存储在数据库中,设计以语义概念为检索条件的用户检索界面,根据置信度的高低排序,返回包含用户感兴趣语义概念的视频镜头,从而实现在语义层对视频内容的检索。

3　系统各组成部分的设计与实现

3.1　视频镜头边界检测与关键帧提取

镜头是视频检索的基本单位,镜头边界通常可以被分为:突变(Cut)、淡入淡出(Fade in/out)和溶解(Dissolve)3 种类型。系统中,将这 3 种不同的镜头边界的检测分为 3 个独立的通道来处理。

对于突变类型的镜头边界,采用比较前后帧直方图的方法,产生镜头边界的候选位置。然后将图像分为 4×4 的块,分别统计每块中的角点数目,如果角点变化的块数大于预先设定的阈值,则认为是突变镜头边界。

对于淡入淡出镜头边界的检测,则统计各候选边界帧的角点总数目,如果等于 0,则认为该帧是淡入淡出镜头边界的中心帧。

溶解镜头边界是一种较为复杂的渐变边界,可通过对视频加一个 10 帧长度的窗来统计连续 10 帧的变化情况。对 10 帧的 4×4 块中每块的角点数目进行最小二乘拟合,如果拟合斜率大于设定阈值,则认为找到了一个候选的渐变边界。再比较第 1 帧和第 10 帧的颜色直方图,如果大于阈值,则认为找到了一个溶解镜头边界。

关键帧提取的目的是为了提取出能有效代表视频镜头语义内容的一幅或者几幅图片。首先统计镜头中各帧图像的直方图(通常是颜色直方图)信息;然后归一化任意两帧图像之间的直方图差;判别帧间是否存在语义差,再建立一个语义差矩阵。语义差矩阵是一个二值的矩阵,它表征了镜头中各帧序列之间是否存在着语义差异。系统中采用动态规划

算法分割该矩阵,这相当于对镜头内的帧序列进行时域上的分割,再寻找最优的分割位置和数量,最后选取每个时域分割段的中间帧作为关键帧。

3.2 语义概念提取

如何有效提取视频中的语义概念是系统性能优劣的关键所在。首先要在关键帧的基础上提取图像低层特征,包括颜色、形状和纹理特征。如何跨越图像低层特征与高层特征(即语义概念)之间的语义鸿沟,一直是国内外基于内容图像检索领域研究的热点和难点。在系统中,使用支持向量机(SVM)来识别图像中包含的高层语义概念。再将关键帧与视频镜头对应起来,即如果某个视频镜头中有任意一个关键帧包含某个高层语义概念,则认为这个视频镜头包含了这个语义概念。

在系统中,共提取了 4 种图像低层特征:

(1)颜色矩特征(CMG);

(2)颜色自动相关图特征(CC);

(3)边缘方向直方图特征(EDH);

(4)共生矩阵特征(CT)。

其中,前两种特征属于颜色特征,后两种分别属形状特征和纹理特征。

系统使用模式识别的方法来进行语义概念提取,其具体流程如图 2 所示。从图 2 中可以看出,对已提取的 4 种低层特征,首先分别构造最优的 SVM 分类器,再将各个最优 SVM 分类器的预测结果进行融合,得到最终的分类置信度结果。

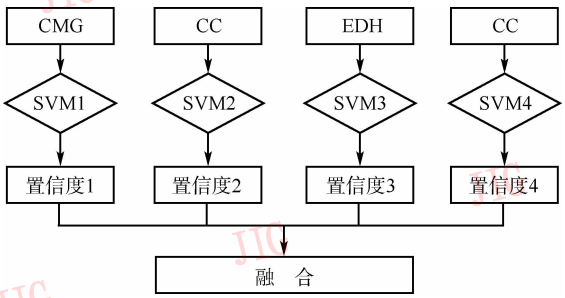

图 2 语义概念检测流程图

Fig. 2 Flow sheet of semantic concept detection

由于 SVM 是一种对核函数类型和参数非常敏感的学习机器,构造最优的 SVM 分类器模型才能保证其识别的准确率。所以,系统选取在图像检索领域广泛使用的 RBF 核函数,它的具体形式如下:

$$K(x,z) = \exp(-\gamma\|x-z\|^2) \qquad \gamma > 0 \quad (1)$$

式中,γ 为核函数参数。

系统中采用交叉验证的方法选取最佳的核参数。为克服训练数据正负样本的不平衡性,需要对负样本数据进行欠采样处理,构建平衡的训练数据集合。

为适应检索的目的,改变了标准的二元 SVM 分类器的输出结果,即不再是“正”或“负”的判别结果,而是使其判别输出包含某个概念的置信度,这个置信度值在 0 ~ 1 之间,代表了测试样本(图像)中包含某个概念的可能性。

对于各个不同的 SVM 分类器的输出结果,采用基于验证平均准确率的线性加权方法进行融合。首先归一化各个不同低层特征分类器的验证平均准确率,得到其投票权重

$$w_j = \frac{VP_j}{\sum_{i=1}^{m} VP_j} \qquad (2)$$

式中,VP_j 代表验证平均准确率。

然后,对各个分类器的输出结果进行线性加权融合,得到最终预测结果

$$\gamma_i^{\mathrm{Comb}} = \sum_{j=1}^{m} w_j \cdot \gamma_i^{(j)} \qquad i = 1,\cdots,n \qquad (3)$$

式中,w_j 和 $\gamma_i^{(j)}$ 分别为不同分类器的加权权重和预测置信度,γ_i^{Comb} 表示融合后的最终分类置信度。

3.3 用户检索

用户检索部分由数据库存储和图形用户界面共同组成。为了实现对语义概念的检索,在得到关键帧语义概念的预测置信度之后,需要将这些信息存储到数据库中。系统采用 SQL 来存储这些数据,并建立了两个数据库表。一个数据表用于存储每个关键帧对应的全部 37 个语义概念的预测置信度值;另一个数据表用于存储每个视频镜头对应的全部关键帧的帧号。

用户界面如图 3 所示。界面的左半部分列出了

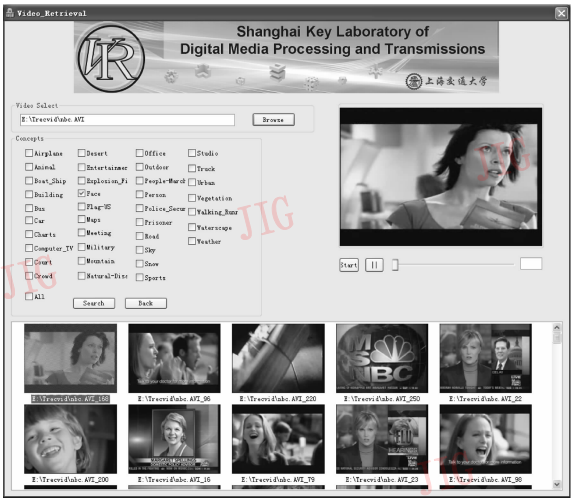

图 3 系统用户界面

Fig. 3 Graphic user interface

37 个可选概念,用户可以选择任意一个概念作为检索条件。系统将返回包含该概念的视频镜头,根据预测置信度的高低对镜头进行排序(即置信度高的镜头排在前面),并显示在用户界面的下方。在返回结果的显示框中,每幅图片都代表一个镜头的关键帧,点击该图片,将在界面的右上部分播放该镜头的内容。

4　系统测试结果及分析

TRECIVD2005 定义了 37 个语义概念,最终选用了其中的 10 个概念作为评测标准,并采用平均准确率(AP)来对检索结果进行评估。平均准确率是一种广泛应用于信息检索领域的评估准则,它既表征了检索结果的准确率,又评估了检索结果的排序相关性,是测试视频检索系统性能的通用评估准则。

为了对系统的性能进行客观地评价,在 TRECVID2005 的视频数据上进行了测试。图 4 显示了系统的平均准确率与 TRECVID2005 的评测结果的对比。从图 4 中可以看出,系统概念检测的平均准确率在 10 个概念上都远高于中等水平,部分概念的平均准确率达到或超过最高水平。10 个概念的平均准确率的统计平均值与 TRECVID2005 的最佳结果相当。

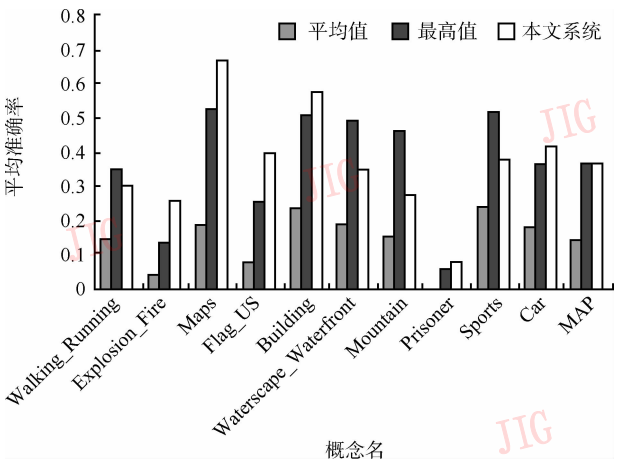

图 4　TRECVID2005 评测概念的 AP 比较

Fig. 4　Compare of AP on different concepts in TRECVID2005

5　结　论

设计并实现了一种基于语义概念的视频检索系统。首先采用视频镜头分割和关键帧提取技术对视频数据进行层次分割,然后对所有关键帧分别提取颜色、纹理和形状特征,再构建出最优模型的 SVM 分类器,挖掘出图像的高层语义概念,并使用基于先验知识的线性加权方法进行融合。系统实现了基于 TRECVID2005 定义的全部 37 个概念的检索功能,并与 TRECVID2005 的 10 个评测概念进行检索性能的比较,实验结果表明,该系统达到了 TRECVID2005 的最佳性能,取得了令人满意的效果。

参考文献(References)

1　Smeaton A F, Over P, Kraaij W. Evaluation Campaigns and TRECVid [DB/OL]. http://doi. acm. org/10. 1145/1178677. 117872, 2006.

2　Cotsaces C, Nikolaidis N, Pitas I. Video shot detection and condensed representation [J]. IEEE signal processing magazine, 2006, **23**(2): 28 ~ 37.

3　Amir A, Argillander J, Campbell M. Ibm Research Trecvid-2005 Video Retrieval System [EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tv. pubs. org. html,2005.

4　Chang S F, Hsu W, Kennedy L. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction [EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tv. pubs. org. html,2005.

5　Chang Chih-chung, Lin Chih-jen. LIBSVM: A Library for Support Vector Machines [CP/OL]. http://www. csie. ntu. edu. tw/ ~ cjlin/libsvm, 2001.