视频相似度的衡量

吴 翌 庄越挺 潘云鹤
(浙江大学人工智能研究所 杭州 310027)
(浙江大学-微软视觉感知实验室 杭州 310027)

摘 要 基于内容的视频检索系统中,最常用的检索方式是例子视频查询,即用户提交一部视频,系统返回相似的一系列视频.但是,怎样定义的两部视频是相似的,仍然是一个困难的问题.文中介绍了一种新的方法以解决这一难点.首先,提出了镜头质心特征向量的概念,减少了关键帧特征的存储量.其次,利用人类视觉判断中所潜在的因子,提出了视频在镜头间相似度的衡量,以及总体上相似度的衡量的方法,为不同粒度上的衡量提供了很大的灵活性,在现实意义上也是合理的.检索实验的结果证明了算法的有效性.

关键词 基于内容的视频检索,例子视频查询,镜头质心特征向量
中图法分类号 TP391.4

Video Similarity Measurement

WU Yi ZHUANG Yue-Ting PAN Yun-He
(*Institute of Artificial Intelligence, Zhejiang University, Hangzhou* 310027)
(*Microsoft Visual Perception Laboratory, Zhejiang University, Hangzhou* 310027)

Abstract The main retrieval method of content-based video retrieval system is query by example. If user submits a video as example, system returns a set of similar videos. But how to define whether two videos are similar is still a great problem. This paper puts forward a video similarity model to solve the difficulty. First, it advances the centroid feature vector of shot in order to reduce the storage of video database. Second, considering the latent factors existing in human's vision perception, it introduces a new comparison algorithm based on multi-level of video structure, such as from shot's view and from the overall view. This different granularity of measurement provides great flexibility, which is reasonable in real world. The final retrieval result demonstrates the validity of algorithm.

Key words content-based video retrieval system, query by example, centroid feature vector of shot

1 引 言

开发研究视频检索系统是当今多媒体应用领域中一个很有发展前景的课题.由于视频所包含的内容极其丰富,难以用文字描述,而且人工注释关键字带来了繁琐性和主观性的弊端,更令研究者意识到在视频检索领域中单纯依靠文本信息作为索引是不现实的.所以,虽然在视频数据库中仍然保留着文本注释的索引,但它不是用来描述视频的具体内容,而是在有限的范围内描述一些确定性特征,比如: 导演、出品日期等.

原稿收到日期: 1999-12-13;修改稿收到日期: 2000-04-27.本课题得到国家自然科学基金(69803009)资助.吴 翌,女,1977年生,硕士研究生,主要研究方向为视频处理、多媒体技术.庄越挺,男,1965年生,博士,副教授,主要研究方向为多媒体数据库、智能 CAD系统.潘云鹤,男,1946年生,中国工程院院士,教授,博士生导师,主要研究方向为计算机美术、形象思维、智能 CAD系统、GIS、计算机动画和多媒体技术等.

基于内容的视频检索主要是依赖它的视觉特征以及时空特性,最常用的检索方式是提交例子视频,查询类似的视频.它与关键字注释等其它方式结合,能在大型视频数据库中检索到用户需要的视频.并实现以下功能:

(1) 高效性.能够快速地找到用户所需要的视频(比如在 WWW 范围).

(2) 简便性.用户操作简单、便捷,不需要掌握复杂的专业领域知识.

(3) 正确性.返回的视频应当尽可能地接近用户的检索要求.

虽然基于内容的视频检索有很多基于文本的检索系统(Yahoo、Sohu等)所无法媲美的优势,但它设计实现的复杂性也随之增加.要想设计出好的基于内容的视频检索系统,必须定义怎样的视频算是相似,解决以下的技术难点:

(1) 视频不是简单的帧序列集合,而是由场景—组—镜头—关键帧组成的层次结构[1].视频间相似度衡量在哪个层次上进行,是视频比较的前提.

(2) 关键帧的视觉特征是整部视频视觉特征的基础,但每部视频都有相当数量的关键帧.对于大型的视频数据库而言,所有视频的各个关键帧视觉特征的存储量和相互间比较次数都是很可观的.

(3) 两部视频是否相似是一个很复杂的问题,不同的人有不同的理解,掺杂着人的主观因素,要设计合理的视频比较算法,必须综合考虑各种因素.

国内外研究者已经在这些方面进行了有意义的研究. Dimitiova, Abdel-mottaled[2]以两个视频间对应帧的平均距离作为相似度,并规定视频帧序列遵守时间顺序. Lienhart等[3]从不同层次考虑了视频的相似度问题,并在表示集与表示序列上以不同聚集度定义了相似度.但他们所考虑的对度量的影响因素十分有限,人们对视觉方面的判断往往有许多标准,而只有当我们的相似度度量能模拟人的判断时才是有意义的.

本文提出的视频比较算法,出色地解决了以上的问题,任意两部视频都可以在有限的时间内得到其相似度.实现了面向 WWW 的基于内容视频检索系统 WebScope-CBVR.

2　系统综述

我们的视频检索系统的结构图如图 1所示,分成 3个模块.

(1) 视频获取模块. 视频的获取是进行视频处理的前提.系统设计了一个 Crawler,从初始网站开始,它根据不同的链接不停地寻找,下载合适的视频到数据库.从视频的 URL和主页的 HTML文件中,尽可能地提取关于该视频的文本信息,如: 视频的题目、摘要、类别、主题、出品日期,以及相关的一些人物信息,如: 导演、制片人、演员等.

(2) 视频处理模块.提取视频的特征.

(a) 从 URL和 HTML中提取的、以及一些人工注释的文本信息,作为视频的属性特征.

(b) 我们采用颜色直方图结合图形学中边检测的方式,提取视频底层特征-颜色分布信息,进行镜头边缘的检测与切分[4];并采用基于非监督聚类的方法提取关键帧[5];同时,在镜头序列的基础上,根据语义信息构造场景和组结构[5].视频就表示成为一个分层次的结构,作为视频的结构特征.

(c) 在视频结构化的基础上,对于结构化之后的每一个关键帧,我们都利用图像处理技术提取其视觉特征,将其结合起来作为视频的总的视觉特征.

(3) 视频的检索.该模块是整个检索系统面向用户的重要部分.用户的检索要求主要是以提交视频作为例子的方式进行的.系统通过视频相似度的衡量算法,检索出最相似的一系列视频,返回给系统.设计好的视频相似度算法是实现视频检索出色效果的前提,本文着重介绍了视频的相似度衡量方法.

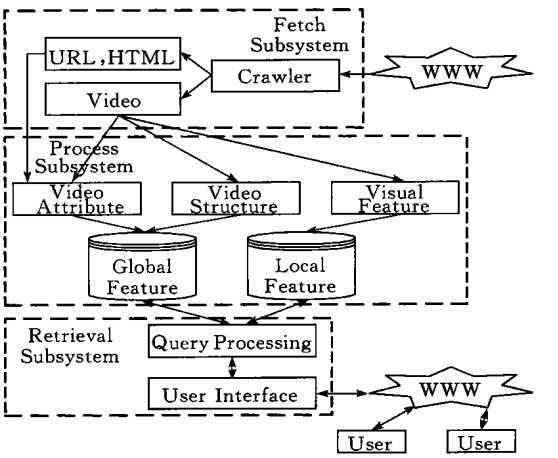

图 1　系统结构图

3　镜头质心特征向量

通过视频结构化工作[4]后就得到了视频的层次结构(见图 2).每个层次均由数个下一层的内容组成,如一个视频由数个场景组成,一个场景由数个组

组成,一个组由数个镜头组成,一个镜头由数个关键帧组成.最底层是关键帧,视频所有关键帧的视觉特征集合代表该视频的视觉特征.一般而言,视频的关键帧数量还是很多的,相应对数据库存储容量要求也很大.

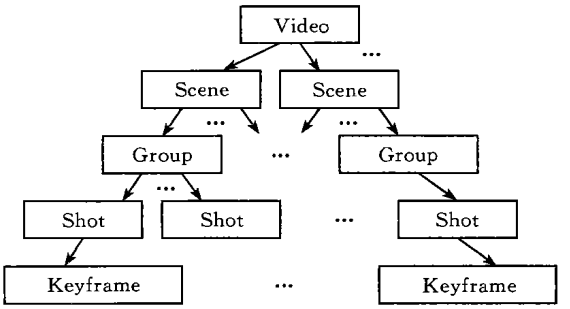

图 2 视频结构图

目前系统中提取的视觉特征有颜色和纹理两类.颜色特征用 HSV 空间直方图来表示的,仅利用 H, S 分量在 8× 4的二维空间中统计直方图,以归一化后的 32个数值作为特征值.我们的纹理特征是用 Tamura 等[6]定义的粗糙度,对比度和方向性这三个数值组成的一个分量来表示.所以,最终每个关键帧的视觉特征表示为 35个数值的多维向量.大量多维向量的比较不仅在时间效率上很低,而且对数据库存储空间耗费也很大.

考虑到镜头内部各关键帧之间的视觉特征差别与不同镜头的关键帧之间的视觉特征差别不算很大,有可利用的冗余性.为了提高空间、时间的效率,我们可以近似地将一个镜头中所有关键帧特征向量用一个质心向量来表示,用以代表整个镜头的视觉特征向量.图 3中四周散乱的小黑点代表各个特征向量,趋向集中于中间的黑点—— 质心向量.

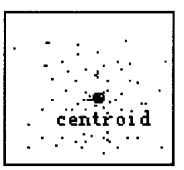

图 3 特征向量与质心向量

求镜头的质心特征向量:

(1) 设镜头 S 由 m 个关键帧组成, $S=\{K_1, K_2, \cdots, K_m\}$, 每个关键帧 K_i 由 35个特征向量表示, $K_1=\{F_{11}, F_{12}, \cdots, F_{1N}\}, K_2=\{F_{21}, F_{22}, \cdots, F_{2N}\}, \cdots, K_m=\{F_{m1}, F_{m2}, \cdots, F_{mN}\}$, $N=35$.

(2) 对每一维 F_{ij}, $j=1, \cdots, 35$,求质心向量每一分量 $\boldsymbol{F}_i$.

$$\boldsymbol{F}_i=\frac{\sqrt{\sum_{j=1}^{N=35}\boldsymbol{F}_{ij}^2}}{N},\quad i=1, \cdots, m \tag{1}$$

(3) 最终镜头由一个 $N=35$维的质心向量 $\bar{\boldsymbol{K}}$ 表示, $S=\{\boldsymbol{K}\}=\{\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_N\}$.

由于同一镜头中各关键帧的视觉特征相差不大,所以最终形成的质心向量的每一分量 $\boldsymbol{F}_i$ 与 F_{ij}, $j=1, \cdots, 35$偏差不会很大.用 $\boldsymbol{K}$ 代替镜头中所有关键帧的视觉特征不会造成太大误差,却大大减少了视频数据库的存储量.

4 基于 Shot层次的比较算法

镜头是视频结构中仅高于视频帧的层次 (见图 2),包含了丰富的视觉和语义内容.视频间相似度的比较建立在镜头间相似性的基础上.基于 Shot层的比较算法,要实现对于提交的例子视频 V_q 中的任意一个镜头 $S_{qi}(i=1, \cdots, n)$,在视频数据库所有视频 V_d 中找出与之相似的镜头.下面我们就考虑任意两个镜头 *Shot* 1与 *Shot* 2之间的相似度度量方法.在第 3节中我们用一个质心特征向量表示一个镜头,所以 *Shot* 1与 *Shot* 2相似度的度量转化为 35维的质心特征向量 $\boldsymbol{K}_1$ 与 $\boldsymbol{K}_2$ 之间的相似度的度量.

32个数值的颜色特征分量,由于它们已经归一到 0与 1之间,并定义在同一物理域中,因此,只要用如下欧拉距离计算颜色特征的相似度:

$$Similarity_C=\sum_{i=0}^{31}\mathrm{Min}(\boldsymbol{K}_{1\text{colorhistogram}}(i), \boldsymbol{K}_{2\text{colorhistogram}}(i)) \tag{2}$$

而对于 3个数值的纹理特征分量,其每个数值都在不同的范围内.因此,我们必须首先将其每个数值都用高斯归一化方法转换到同样的范围中 (如 - 1到 1之间).设视频库中共有 k 个镜头,纹理特征为 $Texture=(Coarse_i, Contrast_i, direction_i)(i=1, \cdots, k)$.我们首先计算这 k 个值的均值 m 与方差 σ,然后将 k 个值都按下式归一化到 - 1与 1之间

$$Texture_i'=\frac{Texture_i-m}{3\sigma} \tag{3}$$

用欧拉距离计算纹理相似度,并用下面的线性变换将其转化到 [0, 1]之间.

$$Similarity_T'=1-\frac{Similarity_T+1}{2} \tag{4}$$

最终,根据颜色、纹理分量的权重 (W_c, W_t)计算 *Shot* 1与 *Shot* 2的相似度.

$$Similarity(\boldsymbol{K}_1, \boldsymbol{K}_2)=W_c\times Similarity_C+W_t\times Similarity_T' \tag{5}$$

但是我们可以看到,以上仅仅考虑了镜头在视觉特征向量上的差异,忽视了各个镜头时间长度的不同.设通过上述质心向量相似度算法式(5)得到了与 S_{qi}最相似的镜头为 S_{dj}, S_{dk};其中, S_{qi}的长度为 30帧, S_{dj}的长度为 50帧, S_{dk}的长度为 32帧.虽然从质心特征向量看 S_{dj}, S_{dk}与 S_{qi}间相似度一样,但是 S_{dk}是 S_{qi}更长的版本,而 S_{dj}与 S_{qi}几乎是相同的, S_{dj}与 S_{qi}相似度更大.所以,我们必须考虑镜头长度这个因素.它可以度量出同一镜头的快慢不同版本的差别程度,体现了人们视觉判断中的时间跨度相似性标准.

$$length = 1 - \frac{|\ length_{qi} - length_{di}\ |}{length_{qi}} \tag{6}$$

综合以上两点,我们用式(7)计算最终的镜头相似度,并转化到 0与 1之间. 1表示最为相似, 0表示完全不相似.

$$S\_\ Similarity(S_1, S_2) = W_v \times Similarity(\boldsymbol{K}_1, \boldsymbol{K}_2) + W_l \times length \tag{7}$$

其中, W_l, W_v分别为 $length$和 $Similarity(\boldsymbol{K}_1, \boldsymbol{K}_2)$的权重.

当然,镜头长度的不同与视觉特征相比,并不是一个最根本的区别依据.但是在视觉效果基本相同的情况下,长度相同的镜头比长度不同的镜头相似程度更大一些,这也是非常合理的.因此,镜头长度这个因子,可以作为一个附加的判断因子;当然 W_l与 W_v相比所占的比例要小很多.

5 视频的总体比较算法

视频由若干个镜头组成,经过第 4节的计算,我们已经得到了 V_q中某个 S_{qi}($i=1,\cdots,n$)与数据库中所有镜头(假设共有 K个)之间相似度的值 $Similarity_i$($i=1,\cdots,K$).那么,在这 K个 0, 1之间的值中,哪些才算相似镜头所具有的值呢?这实际上相当于在 0和 1之间找一个动态阈值 $Threshold$,所有大于 $Threshold$的值我们认为它代表了相似镜头,反之则为不相似.我们应用聚类算法式(8)来寻找 $Threshold$值,以确定所有数据库中的与 S_{qi}相似的镜头的最小相似值.

$$Threshold = \{Similarity_i \mid \mathrm{Max}\left\{\frac{\sum_{j=i+1}^{K} S\_\ Similarity_j}{K-i} - \frac{\sum_{j=1}^{i} S\_\ Similarity_j}{i}\right\},\ i=1,2,\cdots,K\} \tag{8}$$

相对于通常的预定义阈值的方法,这种计算方法具有很强的灵活性.它能根据当前的相似度状况动态地判断相似性.在文献[7]中, S_{qi}分别对各个视频段求最相似的镜头,使得一些 S_{qi}与根本不相似的视频也必须求出相似对应镜头来,很不合理.而我们的方法,在整个视频数据库所有镜头范围内,找与 S_{qi}最相似的若干镜头,如果某 V_d与 V_q的所有对应镜头都为不相似,则该 V_d就不可能是结果返回的视频.这样该对应在以后的计算中就不用考虑了.

在排除了与 V_q没有相似镜头的视频后,得到了若干个视频集合 $V_d\_set = \{V_{di}\}$, S_{qi}($i=1,\cdots,n$)与 $\{V_{di}\}$中所有镜头的对应如图 4,相似镜头用连线表示. V_q的第三个镜头与 V_{d1}的第三个镜头相似,与 V_{d2}的第二个镜头相似.

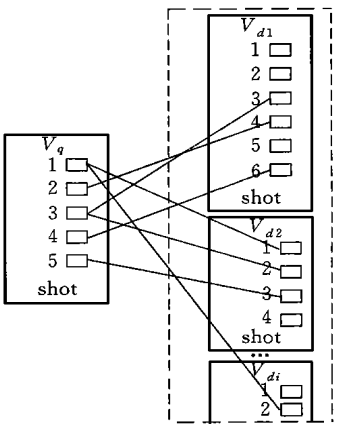

图 4　视频的相似镜头对应图

设 V_q, V_{di}各有 n和 m个镜头,各有 n'与 m'个相似对应镜头.下面要判断 n'和 m'个镜头是否以相同的时间顺序出现.如图 4所示, V_q的第 2, 3镜头与 V_{d1}的对应镜头就出现了逆序. $forward(correspond_shot(i))$记录对应镜头是顺序, $back(correspond_shot(i))$记录对应镜头是逆序.通过式(9)可以计算出 V_q, V_{di}总的顺序、逆序对应数.我们以一个 0, 1之间的数值 $Order$来衡量对应图中出现这种逆序的程度. 1代表两个视频的对应镜头完全是顺序一致的, 0则为完全是逆序.在影片中逆序就可能是后期剪辑产生的效果.这个因子体现了人们视觉判断中的时间顺序相似性标准.

$$Order = \frac{\sum_{i=1}^{n'} forward(correspond\_shot(i))}{\sum_{i=1}^{n} forward(correspond\_shot(i)) + back(correpond\_shot(i))} \tag{9}$$

另外,在 $n > n'$, $m > m'$情况下, V_q与 V_d各有一些零星的镜头找不到对应,如图 4中 V_{d1}的第 1, 2, 5个镜头在 V_q中未找到相似的对应镜头,同样 V_q的第 1个和第 5个镜头在 V_{d1}中也找不到相似镜头.这样的

镜头越多,相似程度越差,体现了对应的不连续性.我们以式(10)度量人们视觉判断中时间连续性的标准.

$$continuity = \frac{n'}{n} \times \frac{m'}{m} \quad (10)$$

在衡量了上述所有因素以后,可以得到视频的总体相似度

$$Similarity = W_1 \times \sum_{i=1}^{n} \left(\frac{length_{qi}}{Totallength_q} \times S\_Similarity_i \right) + W_2 \times Order + W_3 \times continuity \quad (11)$$

其中 W_i 表明了人们对各种因素的重视程度. $S\_Similarity_i$ 之前的分数是该镜头在整个视频中所占时间的百分比,用它作为权重体现了在视频中占较大比例的镜头将对整个视频的相似度有较大的贡献,这符合正常人体感受.当 V_q 与数据库中的任意视频 V_d 的相似度都计算完后,只要排列出相似度最大的若干个 V_d 返回给用户就完成了查询.

6 实验结果及评价

我们在微机上完成了视频检索系统 Webscope-CBVR.该系统由前后台程序分别完成,后台程序 1) Crawler一直在服务器端运行,不停地在 Web上检索相关视频,下载视频至视频数据库, Crawler用 C++ 程序实现. 2)服务器有专门程序负责对视频进行结构化,分析视频的视觉特征,提取特征向量,存储在视频特征数据库中,用 C++ 程序实现.前台程序 3)负责构建客户端界面,通过视频相似度和衡量算法,处理用户的各种查询要求,用 Active Server Page实现.我们的视频数据库采用的是 SQL Server,目前数据库中已有相当数量的视频段,其中包括电影、系列片、电视广告、体育比赛和一些录像片段.

我们把图 5的第一个视频提交给系统作为例子进行查询.结果返回界面如图 5所示,该图中的第一行是我们提交的例子,其余的都是通过相似度算法寻找到的相似视频段.在我们的系统实现中,由该算法完成的检索模块显示了出色的效果.

图 5 以第一个视频为例,检索到的相似视频

7 总 结

视频段的相似度度量是视频检索中的关键问题.相对于已有的算法,本文在充分衡量了视频数据库存储和视频间比较的时空效率后,提出了镜头质心向量的概念,减少了关键帧视觉特征向量的存储.同时,综合考虑人们主观视觉判断因子,提出了全面衡量视频间的相似程度的模型.该算法从视频各个层次结构出发,不但具有粒度性,还具有整体性,可用于不同层次的视频结构的比较.目前,我们正在处理用户的相关性反馈上做进一步的工作.这样,各个权重就能自动地调整,以体现用户对不同因子的偏好,把更好的检索结果返回给用户.

参 考 文 献

1 Madirakshi Das, Shih-Ping Liou. A new hybrid approach to video organization for content-based indexing. In: Proceedings of IEEE Conference on Multimedia Computing and Systems, Austin, Texas, 1998. 96-100

2 Nevenka Dimitrova, Mohamed Abdel-Mottaled. Content-based video retrieval by example video clip. In: Proceedings of the International Society for Optical Engineering, San Diego, 1998, 3022 214-221

3 Rainer Lienhart, Wolfgang Effelsberg, Ramesh Jain. VisualGREP: A systematic method to compare and retrieval video sequences. SPIE, 1997, 3312 271-282

4 Zhuang Yue-Ting, Wu Yi, Pan Yun-He. Video catalog—— A new approach to video organization. Pattern Recognition and Artificial Intelligence, 1999, 12(4): 408-415(in Chinese)
(庄越挺,吴 翌,潘云鹤. 视频目录—— 一视频结构化的新方法. 模式识别与人工智能, 1999, 12(4): 408-415)

5 Yue-Ting Zhuang, Yong Rui, Thomas S Huang. Adaptive key frame extraction using unsupervised clustering. In: Proceedings of the IEEE International Conference on Image Process, Chicago, 1998. 76-81

6 H Tamura, S Mori, T Yamawaki. Texture features corresponding to visual perception. IEEE Transactions on System, Man, and Cyb, 1978, SMC-8(6): 312-320

7 Xiao-Ming Liu, Yue-Ting Zhuang, Yun-He Pan. A new approach to retrieval video by example video clip. In: Proceeding of the 7th ACM International Multimedia Conference, Orlando, Florida, 1999. 435-437