文章编号: 1001- 2486(2007) 05- 0071- 06

一种基于 EDU 模型的新闻视频摘要方法*

谢毓湘, 栾悉道, 吴玲达, 肖 鹏
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 现有的视频摘要技术缺乏一个统一、可扩充的视频摘要模型。针对该缺陷, 提出了实体- 描述- 效用模型(简称 EDU 模型), 该模型从视频实体出发, 经过描述得到效用, 并最终根据效用来生成视频摘要, 该模型具有可扩展性。对 EDU 模型进行了详细阐述, 并根据该模型, 提出了新闻视频故事摘要生成的方法。实验结果表明, 该方法具有令人满意的效果。

关键词: EDU 模型; 视频摘要; 新闻视频; 实体; 描述; 效用
中图分类号: TP391 文献标识码: A

A Method of News Video Summarization Based on EDU Model

XIE Yu-xiang, LUAN Xi-dao, WU Ling-da, XIAO Peng
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The current video summarization techniques are in want of a normal and expandable summarization model. To solve this problem, Entity-Description-Utility (EDU), a general expandable model, is proposed. The model starts from video entity, gets utilities after descriptions, and finally generates video summarization by the utilities. The EDU model is described in detail, and a method of news story summarization based on this model is also produced. The experiment proves the effectiveness of the method.

Key words: EDU model; video summarization; news video; entity; description; utility

随着网络以及多媒体技术的迅速发展, 涌现出了大量的数字视频, 如新闻、广告、监控视频等。如何快速浏览大容量的视频数据, 如何获取和表现视频的内容是亟待解决的问题之一。为了解决这个问题, 出现了视频摘要技术。视频摘要, 即以自动或半自动的方式对视频的结构和内容进行分析, 从原视频中提取出有意义的部分, 并将它们以某种方式进行组合, 形成简洁的能够充分表现视频语义内容的概要。它是对长视频内容的简短总结, 通常用一段静态或者动态的图像序列来表示, 并对原始信息予以保留[1]。

有关视频摘要技术的研究最早可以追溯到 CMU 大学开发的 Informedia 工程[2]。之后, Columbia 大学[3]、Philips 研究院、AT&T 实验室、Intel 公司[4]、IBM Almaden 研究中心、德国 Mannheim 大学[5] 和加州大学 Berkeley 分校等大学或机构都展开了此方面的研究, 开发了多种形式的摘要和生成算法。国内微软亚洲研究院和清华大学合作, 在这方面做了大量的工作, 取得了令人瞩目的成果[6-7]。

目前比较流行的视频摘要主要有以下六种形式: 标题、海报、故事板、缩略视频和多媒体视频摘要。本文重点研究以缩略视频为代表的摘要形式。从视频摘要的生成算法来看, 大致可以分为四类[8]: 基于视频采样的生成方法、基于视觉信息的生成方法、融合多特征的生成方法以及基于视频句法语义的生成方法。目前视频摘要技术已经取得了很大的进展, 但在一些关键的技术上还有待突破。最重要的是: 缺乏对整个摘要过程起指导作用的模型。本文研究的目标就是建立一个普遍适用的视频摘要模型, 并以此模型为指导, 实现新闻视频的故事摘要。

* 收稿日期: 2007- 03- 16
基金项目: 国家自然科学基金资助项目(60473117); 国家 863 高技术基金项目(2006AA01Z319)
作者简介: 谢毓湘(1976—), 女, 讲师, 博士。

1 EDU模型

本文将视频摘要的生成过程抽象化,提出了"实体-描述-效用"(EDU)模型。

定义1 实体(Entity)

实体是指视频中的客观存在,可以是概念的,也可以是物理的。自顶向下,可以把视频文件、新闻故事、场景、镜头乃至帧等都看成实体。不同层次的实体按金字塔形状堆砌在一起,形成了视频的结构。高层的实体由低层的实体组成。实体实际上就是视频的一个子集,可以从集合的角度来描述实体。例如,对于一个包含 N 个镜头的视频片段,其中第 k 个镜头可以表示为

$$Shot_k = \{f_s \in P(f) \mid start(k) \leqslant f_s \leqslant end(k)\},\ k \in [1, N] \tag{1}$$

其中 f_s 表示帧,$P(f)$ 表示全部帧的集合,$start(k)$ 和 $end(k)$ 分别表示镜头 k 的开始帧和结束帧。上式意味着第 k 个镜头可以表示成在镜头开始与结束之间的帧的集合。

定义2 描述(Description)

描述是对实体内容的抽象和概括的表述。与实体所包含的原始信息不同,描述是经过加工和提取的信息,更适合人类理解或者是计算机处理的需要。对于不同层次的实体,应该建立不同的描述。对于同一个实体,可能有多种不同的描述角度,把每个描述角度叫做一个描述子,每个描述子的所有可能出现的实例的总和叫做它的实例集。例如,对于实体 E_k,它的描述可以定义为

$$D_{E_k} = \left\{ d_{k1}, d_{k2}, d_{k3}, \cdots \right\} \tag{2}$$

其中,$d_{k1}, d_{k2}, d_{k3}, \cdots$表示该实体的多个描述子,每个用户都可以根据自己的需要来给实体添加描述子,也可以在某个实例集中添加实例。而且,所有描述都是完全共享的。

定义3 效用(Utility)

效用是实体对用户需求贡献的度量。也就是说某个实体对表现整个视频的内容起了多大的作用,它是由实体本身的特性决定的,可以基于实体的描述来评估该实体的效用。例如,对于一个包含 n 个描述子的实体 E_k,给出如下的效用函数:

$$U_{E_k} = \sum_{i=1}^{n} w_i \cdot \bar{d}_i \tag{3}$$

其中,U_{E_k} 表示该实体的总效用值;$\bar{d}_i$ 表示每个描述子的归一化效用,w_i 表示各个描述子的权重,且满足权值之和为1。效用值的大小决定着某个实体在视频中占的分量,也是决定该实体是否应该保留在摘要中的重要依据。

1.1 EDU模型的数学描述

EDU模型可简单地描述如下:

$$EDU = \{E, D, U, \varphi\} \tag{4}$$

其中 E 表示实体集,D 表示描述集,U 表示效用集,φ 表示它们之间关系的集合。

设 $E_v, E_{st}, E_{sc}, E_{sh}, E_f$ 分别表示视频流实体、故事实体、场景实体、镜头实体以及帧实体的集合,且满足 $E_f \subseteq E_{sh} \subseteq E_{sc} \subseteq E_{st} \subseteq E_v \subseteq E$,则 EDU 模型可用下式来描述:

$$U = \varphi(E_\alpha) = \varphi_3 \cdot \varphi_2 \cdot \varphi_1(E_\alpha) \tag{5}$$

其中,$E_\alpha \subseteq E$,$\alpha \in \{f, sh, sc, st, v\}$。$\varphi_1$, φ_2, φ_3 分别表示从实体到实体、从实体到描述、从描述到效用三类操作关系。由上式可看出,φ 操作可等价为 φ_3, φ_2, φ_1 三类操作的点积。上式可进一步用以下公式来描述:

$$E_\beta = \varphi_1(E_\alpha),\ 其中\ E_\beta = (e_1, e_2, \cdots, e_n)^{\mathrm{T}} \tag{6}$$

该公式反映的是一个从实体到实体的过程,这里不妨将 φ_1 操作理解为对视频的切分或聚类的操作。该式表明,对任意实体 E_α 执行切分操作后,将得到 n 个实体,用 E_β 来表示。举个简单的例子,若采用自顶向下的分割操作,设 E_α 表示视频流实体,E_β 表示故事实体。则上式表示视频流实体经过故事

单元探测, 得到了 n 个故事实体。类似地, 若采用自底向上的操作, 设 E_{α} 表示镜头实体, 经过 φ_1 的聚类操作, 则将生成场景实体 E_{β}。

进一步, 实体到描述的过程可用如下公式表示:

$$D = \varphi_2(E), \text{其中 } D = (d_1, d_2, \cdots, d_n)^{\mathrm{T}} = [\overline{d_{ij}}]_{n * m} \tag{7}$$

经过实体到描述的操作, 可以得到描述集 D。由于每个实体都可由 m 个描述子来描述, 故 n 个实体的描述集可用 $n * m$ 的矩阵来描述。其中 $\overline{d_{ij}}$ 表示对第 i 个实体的第 j 个描述归一化后的效用。

第三步, 根据描述生成效用。如下式所示:

$$U = \varphi_3(D), \text{其中 } U = (u_1, u_2, \cdots, u_n)^{\mathrm{T}} \tag{8}$$

其中 φ_3 表示效用函数, 若取其为简单的加权求和函数, 则有 $u_k = \sum_{j=1}^{m} w_j \cdot \overline{d_{kj}}$, 且满足权值之和为 1。

1.2 EDU 模型的结构

EDU 模型体现的是一种由实体生成描述, 再由描述得到效用函数的摘要思想。如图 1 所示。

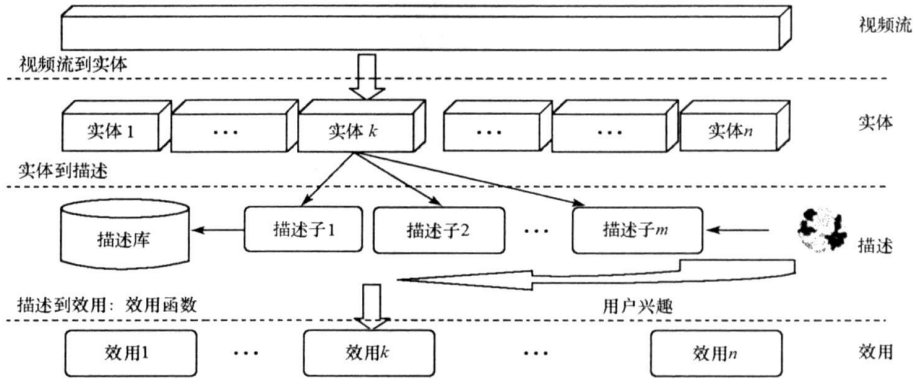

图 1 EDU 模型的结构

Fig. 1 The structure of EDU model

首先, 系统将原始视频流进行切分, 生成不同层次的实体。根据不同的摘要需求, 实体粒度的选择也是不一样的。例如, 要对单个新闻故事进行摘要时, 可以选择镜头作为基本实体; 而若要对整个新闻专题进行摘要, 则选择新闻故事作为基本实体更合适一些。

然后, 采用自动与人工结合的办法对不同层次的实体进行描述。比如, 对镜头实体, 它可能包含多个描述子: 可以用人脸探测算法探测其中是否包含人脸, 如果发现包含人脸, 则为这个镜头实体添加人脸描述子, 并将人脸出现在什么时刻, 出现在画面的什么位置等信息填写到描述子中; 用户还可以手工对镜头进行一些标注, 例如用户觉得某个镜头中包含了暴力冲突场面, 则可以为这个镜头实体的“场面特征”描述子中添加“暴力冲突”信息。本系统提供的是一个开放的描述体系, 每个用户都可以通过算法或人工的方式为实体添加描述子, 描述信息在系统内可以充分的共享。

最后, 在描述信息的基础上, 应用效用函数来计算每个实体的效用值。并由此获得与实体序列相对应的效用值序列, 以此作为生成摘要的数据基础。

从 EDU 模型的结构可以看出, 拥有一个作为中间信息共享平台的描述体系, 以及基于描述体系的效用函数是该模型最大的特点。第一个特点决定了模型具有较好的可扩展性; 而第二个特点则使得用户偏好可以较好地体现在最终的摘要当中。

2 基于 EDU 模型的新闻故事摘要方法

新闻故事是人们理解新闻内容的基本单位, 目前有关新闻视频摘要的研究大部分是针对新闻故事的。对新闻故事的摘要, 其目的就是缩减新闻故事的长度, 给用户提供快速的、一目了然的浏览方式。

由于新闻视频编辑方式的特点,直接把镜头作为新闻故事摘要的基本实体。

2.1 由镜头描述到镜头效用

镜头描述子是生成镜头效用的基本依据,包括镜头类别效用、人脸效用、字幕效用等等。本节首先讨论如何由描述子来计算相应的效用值,然后根据效用函数得出镜头的效用。

2.1.1 镜头类别效用

每个新闻故事都由若干个镜头组成。出现在不同位置的相似镜头对于视频摘要而言是冗余信息,因此必须设法识别出这些相似镜头,然后应用一定的规则选择真正有代表性的部分。为了达到这个目的,首先要将相似的镜头聚合在一起。采用关键帧描述子中记录的关键帧直方图、边缘直方图、主色调直方图等作为镜头特征的代表,应用自校正镜头聚类算法[9]对镜头进行聚类。这种算法从一个初始的分割开始,经多次聚类分裂和合并的迭代,能自动地进行误差校正,聚类效果比较准确。

假设对一则新闻故事获得了 N 个类别,定义其中第 i 个类别的权重 W_i:

$$W_i = \frac{S_i}{\sum_{j=1}^{N} S_j} \tag{9}$$

其中,S_i 表示第 i 个类别的总持续时间。

一般来说,新闻中短而且与其它镜头很相似的镜头不会太重要,而那些出现次数不多,但持续时间很长的镜头一般含有重要信息。也就是说,某个类别的权重越大,其中的镜头的重要度就应该越低。因此,定义第 k 个类别中某个镜头 j 的重要度:

$$I_j = L_j \cdot \log \frac{1}{W_k} \tag{10}$$

其中,L_j 表示第 j 个镜头的持续时间。W_k 表示第 k 个类别的权重。

可以看出,随着类别权重的增大,镜头重要度降低,而随着镜头长度的增加,其重要程度相对上升。然后,将一则新闻故事中重要度最大的镜头的效用设为1,将重要度最小的镜头的效用设为0,并按比例将其它镜头的效用归一化,最终得到镜头类别效用 $\overline{d_1}$。

2.1.2 人脸效用

镜头的人脸描述子不仅包含了人脸图像本身,还包含了人脸出现的时间、人脸的面积、人脸在帧中的相对位置等信息。可以根据人脸的一些物理特征来定义它的效用。一般情况下,视频中大幅的出现在屏幕中央位置的人脸能引起人们较多的关注,这样定义人脸的重要度[6]:

$$I_{face} = \sum_{k=1}^{N} \frac{A_k}{A_{frame}} \times \frac{w_{pos}^{i}}{8} \tag{11}$$

其中,I_{face} 代表该镜头的人脸重要度,N 代表镜头的总帧数,A_k 表示第 k 帧中人脸的面积,A_{frame} 表示帧的面积,$\frac{w_{pos}^{i}}{8}$ 表示位置权重。

由上式可以看出,$I_{face} \in [0, 1)$,而且一般会远远小于1,因为人脸的面积总是只占整个画面面积的很小一部分。类似于镜头类别效用中提到的方法,将人脸重要度最大的镜头的人脸效用值设为1,重要度最小的效用值设为0,其它重要度按比例进行归一化,得到人脸效用 $\overline{d_2}$。

2.1.3 字幕效用

字幕在新闻中出现得相当频繁,是非常重要的语义信息来源和重要度信息来源。经观察发现,在屏幕下方出现带状新闻标题一般是在一则新闻故事开始的时候,如果该新闻以口播帧开始,字幕一般要持续口播的整个过程,还要在进入正式新闻画面后持续一段时间。而对于那些不以口播帧开始的新闻,标题字幕在新闻开始的时候出现,持续若干秒后消失,如果新闻故事比较长,则有可能在一段时间后第二次出现。由此看出,新闻标题字幕是编导人员为了便于观众了解新闻的主题而加入的,是对当前新闻故事的描述,与当前镜头的画面并不一定直接对应。因此,把这种字幕看成是对当前镜头的描述不妥。而对于人物姓名,一般只在人物第一次露面的时候给出,出现的时间较短,但能够准确反映当前画面的内

容, 应该加以积极利用。

根据以上的观察和分析, 决定对字幕采取区别对待的办法(现有的字幕探测算法对常见的新闻节目已经可以区分两类字幕) 。对于新闻标题类字幕, 一般把它作为新闻故事描述的信息源, 不作为镜头效用的一部分, 将其字幕效用设为 0; 对于出现其它有效字幕的镜头, 把字幕效用设为 1, 没有出现字幕的镜头的字幕效用显然也是 0, 以此作为镜头效用$\overline{d_3}$。

2.1.4 其他效用

对于镜头实体而言, 除了上述几种描述子外, 用户可能添加的描述子还有很多。例如, 用户可以为镜头添加运动强度描述子、特殊声响描述子、摄像机运动描述子等, 在这里不一一举例。

2.2 基于可变效用阈值的摘要方法

在对镜头实体进行描述, 计算每个描述子的效用值后, 再根据用户的偏好, 采用效用函数进行计算, 就可以获得每个镜头实体的效用值, 并由此构建了一个与镜头实体序列相对应的效用值序列。效用值与实体一一对应, 以一个大于等于 0 且小于等于 1 的数值来表征该实体对于表现整个新闻故事所起的作用。而实体序列是基于时间轴的序列, 所以, 也可以将效用值序列体现在时间轴上, 如图 2 所示。

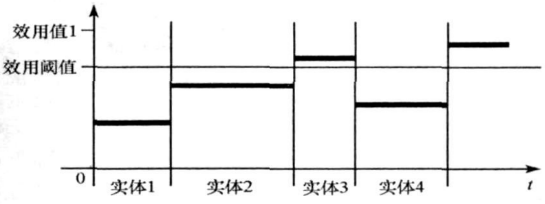

图 2 基于可变效用阈值的摘要方法示意图
Fig. 2 The summarization method based on variable utility threshold

因为一个实体只有唯一的效用值, 而不同的实体的效用值可以各异, 而且每个实体都要持续一段时间, 所以得到的效用值关于时间的函数如图中所示, 是一个阶梯形的函数, 在实体范围内函数值唯一, 而在实体分界处一般不连续。

显然, 效用值大的部分代表的是新闻故事的重要部分, 视频摘要应该从这些部分中选取。图中所示的效用阈值是一个由摘要比率所确定的可变的值。例如, 如果希望得到长度为原视频长度 20% 的缩略视频, 就可以从 1 开始, 逐步降低效用阈值, 将效用大于效用阈值的实体逐一包含到摘要中, 直到摘要长度达到要求为止。

这种算法与原有的基于固定阈值的视频片段取舍方法相比, 更能够准确体现某一片段在整个视频中的地位, 对视频摘要长度的控制也比较精确。

3 实验结果

为评估该模型的效果, 选用最常见的一组描述子和系统默认的用户偏好来设计实验。分别选取了 9′40″的中央电视台晚间新闻、9′55″的凤凰卫视台北直通车和 16′00″的中央电视台世界报道作为实验素材, 按照保留原长度 20% 的压缩比例, 采用三种不同的摘要算法合成缩略视频, 并请了 21 位没有观看过实验素材的同学对摘要结果进行主观评价。

如图 3 所示, 这三种摘要算法分别是: (1) 基于镜头将视音频按比例缩减的方法; (2) 将视频按比例缩减, 并采用口播帧段的语音作为伴音的方法; (3) 基于“实体- 描述- 效用”模型的方法。第一种方法直接从每个镜头的开始部分选取镜头长度的 20% 的视频和音频作为摘要, 第二种方法采用与第一种相同的视频抽取策略, 只是音频采用的是口播帧伴音, 第三种方法就是本系统采用的基于“实体- 描述- 效用”模型的方法, 也采用口播帧的伴音作为音频。

为了防止多次观看内容相同而形式不同的摘要造成理解干扰, 把 21 位同学分为 A、B、C 三组, 用三份素材作三轮测试。每组在一轮测试中只观看一种形式的摘要。比如, 对于晚间新闻, A 组只观看第一种方法生成的摘要, B 组只看第二种, 而 C 组看第三种。每一轮结束后对 A、B、C 三组进行轮换, 让每组都有机会观看各种形式的摘要。观看摘要后, 让被测试者尽自己的理解回答新闻故事中讲述的时间、地点、人物、事件以及摘要的流畅性这五个问题。然后, 再让他们观看原始的新闻视频, 并针对自己刚才对新闻摘要的理解程度打分。例如, 某同学觉得在浏览视频摘要时完全理解了新闻故事的时间, 则可以为

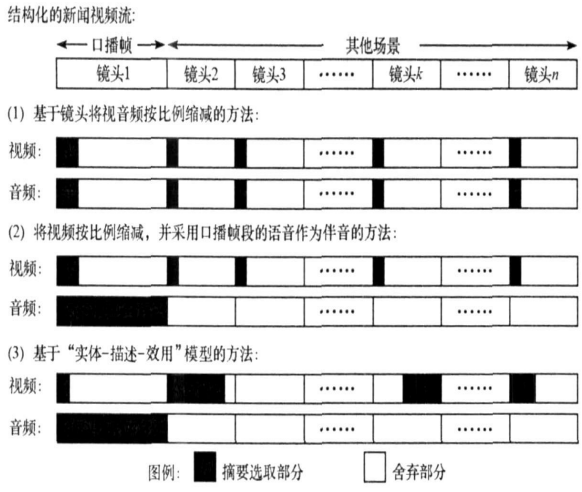

图 3 三种视频摘要方法示意图

Fig. 3 The sketch map of three summarization methods

时间项打满分 10 分,如果认为完全没有获得时间信息,则可以打 0 分,如果认为理解不完全准确,则可以根据自己的理解程度给出 0~ 10 之间的分数。

三轮测试完成后,计算了每种方法下各个问题的平均分,结果如表 1 所示。

由表中结果可以看出,第一种方法由于将视频分解得比较破碎,流畅性非常低,用户对新闻要素的理解程度也比较差;第二种方法由于口播帧语音的作用,用户的理解度有很大提高,但由于视觉感观依然混乱,流畅性还是不高;第三种即本系统的方法提取了效用值高的部分作为摘要,可理解性和流畅性都比前两种有较大提高。

表 1 测试结果统计

Tab. 1 Experiment results

	时间	地点	人物	事件	流畅性
方法 1	5.0	4.0	6.5	2.4	2.4
方法 2	8.2	8.5	7.2	5.6	5.0
方法 3	8.9	9.6	8.9	7.5	7.8

4 结束语

本文提出了一个统一、可扩充的视频摘要模型:"实体- 描述- 效用"模型。该模型从视频实体出发,经过对实体的描述,生成效用值,并根据效用值来生成最终的视频摘要。该模型一方面具有可扩充性,用户可根据自身需要添加不同的描述子;另一方面能够较好地反映用户的兴趣,不同描述子的权值可由用户来指定。根据该模型,可以实现不同层次、不同粒度的视频摘要。以新闻故事的摘要为例,对该模型进行了具体的实验验证。结果表明,以此模型为指导生成的视频摘要能够比较准确地反映新闻视频的重要信息,用户满意程度较高。现有的效用函数采取了简单的加权求和的方法,下一步打算进一步优化效用函数,并将之应用于其他类型的视频摘要中。

参 考 文 献:

[1] Li Y, Zhang T, Tretter D. An Overview of Video Abstraction Techniques[R]. Image Systems Laboratory, HP Laboratory Palo Alto, HPL- 2001- 191, July 31st, 2001.
[2] Christel M G, Smith M A, Taylor C R, et al. Evolving Video Skims into Useful Multimedia Abstractions[C]// Proc. of Conference on Human Factors in Computing Systems (CHI98), 1998: 171- 178.
[3] Sundaram H, Xie L, Chang S F. A Utility Framework for the Automatic Generation of Audio-visual Skims[C]// ACM Multimedia' 02, Dec. 1- 6, 2002, Juan-les-pins, France.
[4] Lienhart R. Dynamic Video Summarization of home Video[C]// Proc. of IS&T/SPIE, 2000, 3972: 378- 389.
[5] Lienhart R, Pfeiffer S, Effelsberg W. Video Abstracting[J]. Communications of the ACM, 1997: 55- 62.
[6] Ma Y F, Lu L. Zhang H J, et al. A User Attention Model for Video Summarization[C]// In Proc. of ACM Multimedia' 02, December, 2002, Juan-les-Pins, France.
[7] 姜帆, 章毓晋. 新闻视频的场景分段索引及摘要生成[J]. 计算机学报, 2003, 26(7): 859- 865.
[8] Uchihashi S, Foote J, Girgensohn A, et al. Boreczky, Video Manga: Generating Semantically Meaningful Video Summaries[C]// ACM Multimedia' 99, 1999.
[9] 熊华, 胡晓峰, 老松杨. 一种自动镜头聚类方法[J]. 国防科技大学学报, 2000, 22(5): 103- 108.