一种草图形式的视频摘要生成方法

陈 佳[1),2)] 滕东兴[1)] 杨海燕[1),2)] 马翠霞[1)] 王宏安[1)]

[1)](中国科学院软件研究所人机交互技术与智能信息处理实验室, 北京 100190) [2)](中国科学院研究生院, 北京 100190)

摘 要 视频摘要作为一种视频内容的简要表示,能够有效地增强用户浏览和组织视频的效率。提出了一种基于草图的视频摘要生成方法。与以往的静态视频摘要方法不同,该方法结合视频内容分析,利用草图在表达上的简洁性和抽象性,对视频中的主要内容进行表达。首先通过视频分析获取视频中的语义特征并提取关键帧,然后通过交互式的方法从关键帧中生成草图,最后进行摘要布局生成完整的视频摘要。实验结果表明,该方法能够有效突出视频的主要对象和主要事件,并具有较高的用户满意度。

关键词 视频摘要 草图 非真实感图形学

中图法分类号: TP301.6 **文献标志码**: A **文章编号**: 1006-8961(2010)08-1139-06

A Novel Method for Generating Sketch form Video Summarization

CHEN Jia[1),2)], TENG Dongxing[1)], YANG Haiyan[1),2)], MA Cuixia[1)], WANG Hongan[1)]

[1)](*Intelligence Engineering Lab.*, *Insititute of Software*, *Chinese Academy of Sciences*, *Beijing* 100190)

[2)](*Graduate University of Chinese Academy of Sciences*, *Beijing* 100190)

Abstract As a brief representation of video content, video summarization can effectively assist users' browsing and organizing video clips. This paper presents a novel method to generate video summarization based on sketches. Based on the result of video content analysis, this method visualizes video's main content, taking advantages of sketches' efficiency and succinctness. Firstly, we extract semantic features and key frames from the video. Secondly, user interaction is adopted to generate sketches from the key frames. Finally, the layout is arranged to get the final summarization result. Experiment result shows that the method can represent the main objects and main events of video materials, and has a quite high user satisfaction.

Keywords video summarization, sketch, non-photorealistic rendering

0 引 言

视频摘要是对视频内容的简短总结,是对视频语义的高度概括和描述。通常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,形成简洁而又能够充分表现视频主要语义内容的概要表示[1]。在视频的浏览和检索过程中,生成有效的视频摘要,有利于减少网络视频传输量,降低视频浏览的时间成本。视频摘要使得用户可以在浏览小数据量内容的情况下对视频有一个基本的了解,因而对于视频传输、媒体资源管理都具有重要意义。

目前的视频摘要技术主要有两种模式:动态视频摘要和静态视频摘要。动态视频摘要可以从较长的视频中生成较短的视频片段,并使用该片段表现较长时长的视频内容,如 Nam 等人[2]通过检测视频中变化显著的片段和语义上的重要信息,采用自适应的视频非线性采样方法, 提出了基于事件的动态摘要框架。与此不同,静态摘要技术采用静态图像的方式来

基金项目:国家自然科学基金项目(60703079);国家重点基础研究发展计划(973)项目(2006CB303105);国家高技术研究发展计划(863)项目(2007AA04Z113)

收稿日期:2010-03-20;**改回日期**:2010-05-10

第一作者简介:陈 佳(1984—),男。中国科学院软件研究所计算机应用技术专业硕士研究生。主要研究方向为人机交互技术。E-mail: kainster1015@gmail.com

表现视频的主要内容,Ma 等人[3]提出利用用户关注模型来获取视频各个帧中的主要区域并将其组合成为视频摘要, Liu 等人[4]提出了 VideoCollage 的方法,从视频中抽取最能表现视频主要内容的帧及图像区域,并根据感兴趣区域(ROI)的显著程度将各个图像区域拼贴到同一张 VideoCollage 图中。Uchihashi[5]选择关键帧,并根据关键帧所描述内容的重要程度确定关键帧的位置和大小,最终生成以连环画的形式作为视频的摘要。它将连续的帧图像通过空间坐标变换无缝拼接在同一画面上,自动拼接成一幅完整的图像来摘要视频的主要内容。

1 系统概要

由于当前的视频摘要方法通常采用视频中原有的帧图像,因此其生成效果容易受到原有帧图像质量的局限:1) 往往含有较多冗余信息,不能够突出视频中的主要对象;2) 缺乏有效表现视频中事件的方法。本文采用草图作为视频摘要的表现方式,因为草图作为一种简洁的图形表示,能够强调图像中的重要信息,而忽略掉图像中的冗余部分,并且草图中抽象的、模糊的形象化信息,可以有效地描述用户意图,表述或增强视频语义,缩小视频低层物理特征与高层语义之间的鸿沟。

本文通过一幅或几幅静态的草图,采用自动生成与用户交互相结合的方式,对一段时间内视频动态的语义信息进行描述。如图 1 所示,系统通过视频内容分析进行镜头分割并获取视频上下文知识,从分割的镜头中提取关键视频帧并进行筛选,通过交互式草图生成的方法生成对于单个视频帧的草图表达,最后利用视频上下文知识对各帧的草图表示进行布局。其生成结果可用于视频海报、视频缩略图,也可以扩展成为视频浏览与组织的新式界面。

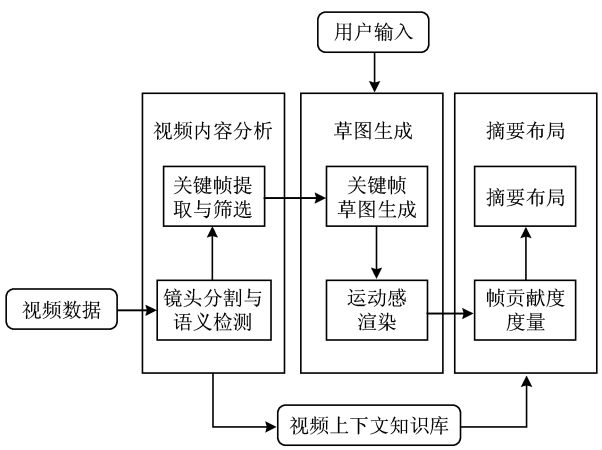


图 1 系统结构图

Fig. 1 System structure

2 视频内容分析

2.1 镜头分割与重要程度评估

镜头是指在一段连续的时间内摄取的一段连续的画面,在每段视频中都含有多个镜头,而这些镜头内包含有不同的语义信息,在整个视频中所起的重要程度也有所不同。依据镜头重要程度进行镜头筛选,以保证视频中重要的语义信息能够在最终生成的视频摘要中得以体现。在使用镜头边界检测的方法将视频分割为多个镜头之后,各个镜头的重要程度 w_{shot} 可以表示为

$$w_{\text{shot}} = \mu T + V \qquad \mu > 0 \tag{1}$$

其中,μ 为经验常数,T 为该镜头中含有的图像帧数,镜头越长,通常其所蕴含信息的重要程度越高;V 为镜头内部各帧的灰度直方图均方差,该值越大说明镜头内部灰度直方图变化越大,发生的事件相应也就越多。通过对每个镜头求 w_{shot} 值,可以得到镜头的重要程度排序。当用户指定所需镜头数目 n 时,则从视频所有镜头中选择值最高的 n 个镜头。

2.2 关键帧图像的提取和筛选

由于镜头本质上仍然是短的视频片段,具有动态的特性,不能够直接转化为静态的摘要形式,因此还需要从镜头内的多幅帧图像中筛选出镜头的关键帧来作为草图生成的初始图像。帧图像的评估值定义为

$$s_{\text{frame}} = Q_t + \eta F \tag{2}$$

其中,F 代表帧图像中检测到的人脸数目。由于大多数视频的主要对象是人物,因此假定出现人脸的图像对于视频的内容表达意义更大,而在存在人脸的帧图像中人脸数目多的又更能完整的表达该镜头内容。帧图像本身的图像质量 Q_t,使用从帧图像中提取到的 SIFT 特征点个数来评估。SIFT 特征对于图像明显变化的区域要比那些模糊的区域更为敏感,因此图像质量定义为 $Q_t = \dfrac{\max N_f - N_t}{\max N_f - \min N_f}$。其中 N_t 代表的是第 t 帧中所提取到的 SIFT 特征点个数,$\max N_f$ 代表所有帧的最高 N_f 值,$\min N_f$ 代表所有帧的最低 N_f 值。

图 2 展示了从一段时长为 5 min,帧速率为 24 帧/s 的视频片段中自动提取出的关键帧图像,每幅图像可以作为一个镜头的表示。为了减少处理时间,在

图 2　从一段时长为 5 min 的视频片段中自动筛选出的关键帧图像

Fig. 2　Key frame images automatically selected from a 5 minutes video segment

进行 SIFT 特征提取前将帧图像重采样为 480 × 360 像素的图像。

3　视频内容的草图表示

3.1　基于 Livewire 的交互式草图生成方法

目前,很多基于边缘检测的方法通过图像处理可以将图像转化为草图风格,如 Zhou 等人[6]提出了一种基于梯度估计的方法自动生成铅笔草图的效果。Fang 等人[7]则利用图像中的颜色分布,提出了一种基于颜色区域的非真实感渲染的方法将彩色图像转化为彩色草图。但是由于图像的复杂性和图像的质量限制,这些基于边缘检测和颜色区域的方法往往不能有效地生成易于理解的草图:一方面这些基于边缘检测的方法对光照强度很敏感,通常只对每一类特定的图像有效,很难对不同光照情况下的图像进行有效的边缘检测;另一方面,即使这些方法能够有效提取出图片中的边缘,现有方法仍然不能对于图像的物体进行区分,不能够判断哪些部分是用户在观看时所真正关注的,在生成的草图中往往包含过多冗余细节,如图 3(b)中图像的主要对象是汽车以及猫,而边缘检测的方法则会将树木的阴影等次要的对象全部提取出来。

本文改进了 Falcao 等人[8]提出的 Livewire 图像分割方法,对于用户指定的起始点和结束点,用动态规划方法寻找一条两点间的最优路径,使该路径的累积代价函数值为极小值。对于路径中相邻两点 p_k 和 p_{k+1},对于 $p_k(x_k,y_k)$,若取搜索窗口宽度为 w,在 $x_k-\frac{w}{2}\leqslant x\leqslant x_k+\frac{w}{2}, y_k-\frac{w}{2}\leqslant y\leqslant y_k+\frac{w}{2}$ 范围内查找 p_{k+1},使得

$$g(p_k,p_{k+1})=\min_{p_{k+1}}\{g(p_{k-1},p_k)+c(p_{k-1},p_k,p_{k+1})\} \tag{3}$$

其中,$g(p_0,p_1)$ 值为 0,且代价函数 $c(p_{k-1},p_k,p_{k+1})$ 定义为

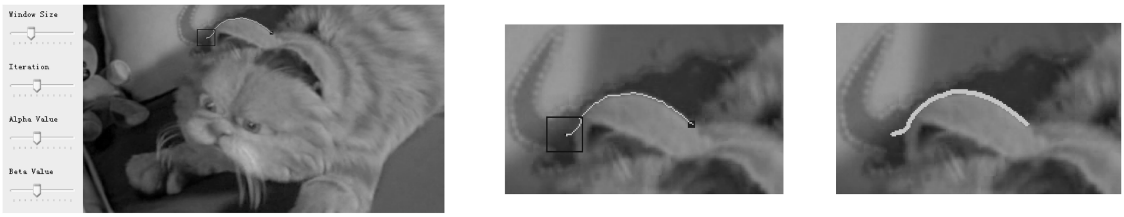
$$c(p_{k-1},p_k,p_{k+1})=\alpha\,|p_{k+1}-p_k|^2+\beta\,|p_{k+1}-2p_k+p_{k-1}|^2+\gamma c_{\mathrm{gra}}(p_k,p_{k+1}) \tag{4}$$

其中,α,β 和 γ 为加权系数,$|p_{k+1}-p_k|^2$ 项为两点空间距离,用于控制笔迹的连续性,$|p_{k+1}-2p_k+p_{k-1}|^2$ 用于控制笔迹的曲率,$c_{\mathrm{gra}}(p_k,p_{k+1})$ 则代表两点间梯度值,使得笔迹能够与图像内对象的边缘相吻合。由于在实际操作中,通过 Livewire 方法提取出的轮廓往往不够平滑,因此用 B 样条曲线对其进行拟合以达到平滑处理的效果。该草图生成方法的用户界面如图4所示,黑色矩形代表路径查找窗口

(a) 原始帧图像　　(b) Canny 边缘检测生成的草图　　(c) 本文中采用交互式方法生成的草图

图 3　基于边缘检测的草图生成与交互式草图生成方法

Fig. 3　Edge detection based sketch generation and interactive sketch generation applied in this paper



(a) 用户界面 (b) Livewire 笔迹 (c) 平滑结果

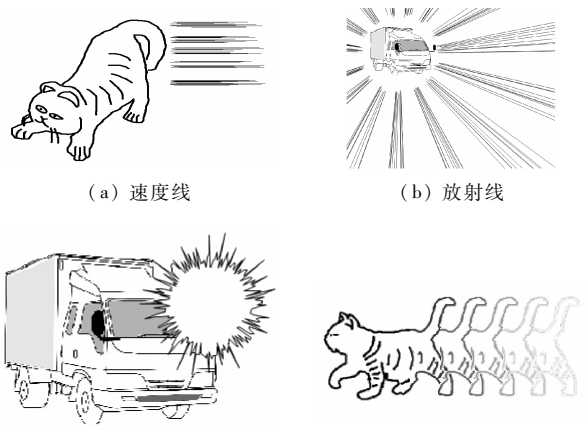
图 4 交互式草图生成界面

Fig. 4 User interface of interactive sketch generation

的大小。用户不需要精确地沿着对象边缘进行绘制,就可以得到平滑的对象边缘。

3.2 面向草图的视频运动感渲染

在采用静态视频摘要表现动态的视频内容时,往往会出现静态的图像不能准确表达视频对象的运动信息的问题。在艺术领域,画家和摄影家会使用多种多样的艺术表现形式来表现运动。而在认知心理学研究中,心理学家也对运动表达方式与人类认知效率之间的关系进行了研究,Friedman 等人[9]通过针对学龄前儿童的实验证明了采用有效的表达方式能够提高人类对于物体运动状态的认知,而 Cutting[10]则归纳出 5 类在静态图片中表现运动的方法,并对其各自的优缺点进行了分析。本文结合草图生成的特点,使用了运动模糊、速度线、放射线、爆炸线 4 种方式来表达物体的运动信息,如图 5 所示,4 种方式适应于不同的情况,并可以多种组合使用。



(a) 速度线 (b) 放射线

(c) 爆炸线 (d) 运动模糊

图 5 运动渲染效果

Fig. 5 Examples of motion rendering

1) 速度线

速度线是在强调物体运动方向时最常用的一种表现运动的方法,可以用箭头、直线、曲线等多种风格表示。本文使用了两种速度线:一是背景速度线,用来强调物体相对于背景的运动,此时速度线覆盖于整个背景中;二是对象速度线,附加于运动对象之后,强调物体的运动方向。在两种速度线中,都使用线条的密度表现对象运动的速度。

2) 放射线

放射线可以产生立体感,其密度可以表现运动物体的运动速度,而放射线的扭曲在表现物体曲线运动时效果通常优于其他方法。

3) 爆炸线

爆炸线通常用来表示物体之间的碰撞,或突发性事件,并可以用来引起观看者的注意。

4) 运动模糊

运动模糊是一种在绘画和摄影中常用的表现方式,不仅能够表现对象运动的方向,还能够体现出对象运动的方式,如平移、旋转、膨胀、缩小等。

本文采用草图用户界面来实现对于可视效果的上述调整,如图 6 中的运动渲染效果的草图用户界面,通过识别用户输入的草图向摘要中添加运动渲染效果,并通过控制线条密度、线长、分组数、内径、外径等参数对其视觉效果进行控制。

图 6 运动渲染的草图用户界面

Fig. 6 Sketch based interface for motion rendering

4 基于视频上下文知识的摘要布局

在进行摘要布局时,首先根据摘要绘图区域大小以及帧图像表达内容的重要程度重新设定帧图像的大小,然后利用视频内容分析得到的镜头间关系

等信息确定帧图像所在位置,最后生成帧图像之间的连接线。

4.1　帧图像大小的确定

在将多个帧图像生成的草图合并到同一幅摘要中时,需要对草图进行缩放。如果缩放比例过小,则草图占整个摘要绘图区域面积比例过小,会出现摘要空白部分过多的现象;如果缩放比例过大,则摘要无法容纳全部草图,会出现帧之间互相遮挡、草图超出摘要边界等现象。同时,考虑到不同帧草图在整个视频摘要所占的重要程度不同,帧草图缩放比例的大小也应当有所不同。该比例可由下式确定:

$$r_i = \alpha \sqrt{\frac{WH}{\sum_{i=1}^{n} w_i h_i} v_i} \tag{5}$$

其中,α 为一经验常数,文中所采用的值为 0.5,W,H 分别代表整个摘要绘图区域的宽度、高度,w_i,h_i 代表该帧图像的宽度、高度,v_i 代表该帧对于整个摘要的贡献值,文中视频内容分析阶段获得的 SIFT 特征点数目来表示。

4.2　摘要位置布局

在确定各个帧图像在摘要中所处位置时,综合考虑视频帧的时间顺序、帧与帧之间的相似关系、连接线出现的交叉点数目等因素,通过动态规划的方法求解出使得目标函数最小的布局来作为最优布局,即

$$Arr(pos_1,\cdots,pos_n) = \arg\min(m_{\text{seq}} + \varphi m_{\text{rel}} + \gamma m_{\text{cross}}) \tag{6}$$

其中,pos_i 代表第 i 幅草图图像的中心点在整个摘要中的坐标,目标函数由 3 部分组成:

1) m_{seq}为时间序列不规则度,假定在布局中时间序列中靠前的帧所在位置总是要比时间上靠后的帧所在位置靠左和靠上,那可以用不符合此规律的帧数目来作为时间序列不规则度的度量

$$m_{\text{seq}} = \sum_{i=2}^{n} \delta x_i + \sum_{i=2}^{n} \delta y_i \tag{7}$$

其中,$\delta x_i = \begin{cases} 0 & x_i > x_{i-1} \\ 1 & \text{其他} \end{cases}$,$\delta y_i = \begin{cases} 0 & y_i > y_{i-1} \\ 1 & \text{其他} \end{cases}$;

2) φm_{rel},m_{rel}代表镜头不相似度度量,文中 m_{rel} 取值为两个关键帧之间的 SIFT 匹配特征点数目的倒数,φ 为经验常数,这里取值为 0.35;

3) γm_{cross},m_{cross}代表连接线交叉程度度量,这里用连接线间交叉点的数目表示,γ 为经验常数,这里取值为 0.25。

通过动态规划方法求式(6)的极小值,可以确定各帧草图位置。该位置可以通过用户输入并进行调整,同时,为了表现视频事件的次序及提高摘要的美观度,使用三次曲线生成不同帧图像之间的连接线。图 7 为系统生成的摘要结果。

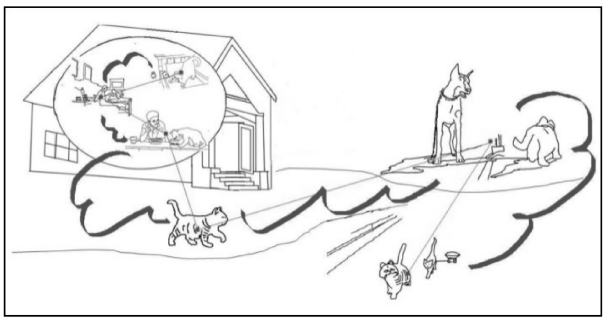


图 7　系统生成结果

Fig. 7　Result of the proposed method

5　实验结果比较与评估

为了对本文方法进行评估,将生成的视频摘要结果与 Image Grabber 视频预览图生成工具以及文献[4]中的 VideoCollage 进行了实验比较。邀请了 12 名被试者参加评估。首先,向每一位被试者呈现一段视频及针对该视频的 3 种摘要:Image Grabber,Video Collage 以及本文的摘要结果,然后,对每一位被试者进行问卷调查,问卷包括 7 个评估维度,评分范围为 1 ~ 6 分。实验数据如表 1 所示。

表 1　实验数据

Tab. 1　Experiment data

问　题	Image Grabber	Video Collage	本文
突出主要对象	2.6	4.6	5.4
突出主要事件	4.3	2.1	5.2
内容覆盖率	4.8	3.2	4.9
美观程度	4.8	5.6	4.9
简洁度	4.8	4.6	4.9
总体满意度	4.6	4.8	5.1

实验结果表明基于草图的视频摘要在突出视频主要对象、主要事件以及内容覆盖率,简洁度方面均优于 Image Grabber 以及 Video Collage,并达到了较高的用户总体满意度。除上述评估结果,被试者指出该方法对于情节性强的视频片段具有更好的摘要效果。同时,部分被试者指出了该方法的不足,如生

成的草图摘要缺少视频原有的颜色信息,影响整体的美观程度等。

6 结 论

针对现有视频摘要方法在表现视频主要信息和视频事件中存在的不足,设计并实现了一种基于草图的视频摘要生成方法。针对其中的关键问题,改进了 Livewire 图像分割算法并将其应用于交互式草图生成,并提出了基于视频上下文知识的摘要布局算法。实验评估表明该方法在表达视频事件和主要内容上优于已有方法,并且具有较高的用户满意度。在下一步工作中,将尝试利用视频图像中的颜色辅助草图生成,并利用草图摘要结果来辅助对于视频片段的检索。

参考文献(References)

[1] Zhu Zhihui. Research on video abstraction techniques[J]. Microelectronics & Computer, 2006, 23(2): 76-82. [朱志辉. 基于视频摘要生成技术的研究[J]. 微电子学与计算机, 2006, 23(2): 76-82.]

[2] Nam J, Tewfik A H. Video abstract of video[C]//Proceedings of IEEE Third Workshop on Multimedia Signal Processing. Copenhagen: Technical University of Denmark, 1999: 117-122.

[3] Ma Y, Hua X, Lu L, et al. A generic framework of user attention model and its application in video summarization[J]. IEEE Transactions on Multimedia Journal, 2005, 7(5): 907-919.

[4] Liu X, Mei T, Hua X, et al. Video collage[C]//Proceedings of the 15th International Conference on Multimedia. Augsburg Germany: ACM, 2007: 461-462.

[5] Uchihashi S, Foote J, Girgensohn A, et al. Video Manga: Generating semantically meaningful video summaries[C]//Proceedings of ACM International Conference on Multimedia. Orlando Florida: ACM, 1999: 383-392.

[6] Zhou J, Li B. Automatic generation of pencil-sketch like drawings from personal photos[C]//Proceedings of IEEE International Conference on Multimedia & Expo. Amsterdam: Institute of Electrical & Electronics Enginee, 2005: 1026-1029.

[7] Wen F, Luan Q, Liang L, et al. Color sketch generation[C]//Proceedings of the 4th international Symposium on Non-Photorealistic Animation and Rendering. New York: Alphascript Publishing, 2006: 47-54.

[8] Falcao A X, Udupa J K, Samarasekera S, et al. User-steered image segmentation paradigms: Live wire and livelane[J]. Graphical Models and Image Processing, 1998, 60(4): 233-260.

[9] Friedman S L, Stevenson M B. Perception of Movement in Pictures[M]//The Perception of Pictures. New York: Academic Press, 1980, 1: 225-255.

[10] Cutting J E. Representing motion in a static image: constraints and parallels in art, science, and popular culture[J]. Perception, 2002, 31(10): 1165-1193.