第36卷 第15期 计 算 机 工 程 2010年8月
Vol.36 No.15 Computer Engineering August 2010
·多媒体技术及应用· 文章编号：1000—3428(2010)15—0211—03 文献标识码：A 中图分类号：TN919.81

基于剧本及字幕信息的视频分割方法

李松斌[1,2]，王玲芳[2]，王劲林[2]
(1. 中国科学院研究生院，北京 100049；2. 中国科学院声学研究所国家网络新媒体工程研究中心，北京 100190)

摘 要：对视频进行分割并建立索引有利于视频的浏览和检索。提出一种融合剧本和字幕对电影类视频进行场景分割与索引的方法。该方法通过字幕与场景中对白文本的匹配获得场景视频片段的初始分割边界，分析视频帧之间的颜色分布差异，以确定场景间的最终边界。实验结果表明，与完全依赖视频内容分析的方法相比，该方法对电影视频场景片段的分割速度和准确度较高。

关键词：视频分割；剧本；字幕；视频内容分析

Video Segmentation Method Based on Film Script and Subtitle Information

LI Song-bin[1,2], WANG Ling-fang[2], WANG Jin-lin[2]
(1. Graduate University of Chinese Academy of Sciences, Beijing 100049;
2. National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190)

【Abstract】Video segmentation and video index favour video browsing and retrieval. This paper proposes a video segmentation and index method based on integration of the film script and subtitles. The method obtains the initial video clips' boundary of different scenes of the video through matching the subtitles and the dialogue text of the scene in the film script. It determines the final boundary of different scenes of the video by analyzing the color distribution difference between video frames. Experimental result shows that the method enhances the video segmentation precision and speed compared with the methods completely depending on content-based analyzing measure.

【Key words】video segmentation; film script; subtitle; video content analysis

1 概述

与镜头检测相比，场景的检测难度更大。目前较通用的视频场景分割方法[1-4]包括 2 个步骤：(1)根据视频图像的颜色、纹理等特征进行镜头边缘检测；(2)通过聚类算法将视觉内容近似且连续的多个镜头划分为一个组并作为一个场景。该方法的缺陷是镜头检测容易受到干扰，例如，对渐变镜头有较高的误判率，而镜头视觉内容的表示和镜头之间相似性的衡量方法目前也不够成熟。上述缺陷导致场景抽取的查全率和查准率不理想。一些针对特定领域视频(如体育比赛和新闻等)的场景分割方法[5-6]一般基于特定应用模型，具有较高的分割精确度，但存在通用性差的问题，即必须针对每个应用事先构造一个视频模型，而模型的构造需要丰富的领域知识。采用此类方法抽取场景后，为支持高效的检索，必须建立说明视频场景内容的高层文本语义索引。

对于电影类视频，剧本是视频内容的天然语义说明，因为视频影像是根据剧本进行拍摄的。剧本一般以场景为故事单元进行组织，剧本中的每个场景一般包含以下信息：地点，时间，内外景说明，场景描述，对白等。如果能将视频按剧本中包含的场景进行分段，那么剧本中场景对应的文本信息可转化为视频场景片段的注释信息，可以避免繁琐的标注工作，且通过结合字幕有利于提高视频场景片段分割的精确度。字幕是场景中的对白，它是视频的部分语义信息，且由于字幕的文本内容与剧本场景中的对白存在对应关系，因此可由字幕获得剧本中每个场景所包含对白在视频片段中的出现和消失时刻，通过综合每个场景所包含对白的出现和消失时刻能大致确定场景所对应视频片段的起始位置。在此基础上通过对视频内容进行分析，可确定场景切换的准确时间点，达到视频的无缝分割。基于上述思想，本文给出一种融合剧本和字幕以及视频内容分析的电影场景视频片段的分割和索引方法。

2 剧本信息的抽取与结构化表示

剧本是影视作品拍摄制作的依据，它是视频(连续图像)的语义信息。一部影视作品的剧本一般由多个场景构成，剧本中的每个场景一般包含地点、时间、内外景说明、场景描述、对白等内容。根据一些正规出版的电影剧本如《中国电影剧本选集》和《谢晋电影选集》等，中文剧本中的场景描述文本一般具有图 1 所示格式。场景表示为一段文本，却蕴涵了电影拍摄所需的各方面信息。剧本通过一些显著的文本编辑特征来区分不同信息，例如，场景基本说明行(首行)一般顶格且为黑体，对白文本的缩进字数要远大于注释文本，对白中说话人用黑体显示等。此类格式信息是进行文本内部表示转换时的重要依据。为便于剧本信息的抽取，本文先根据剧本场景文本的编辑特征将原始场景文本转化为一种内部表示。

基金项目：国家“863”计划基金资助项目(2008AA01Z147)
作者简介：李松斌(1981－)，男，博士研究生，主研方向：多媒体信息处理及检索；王玲芳，副研究员、博士后；王劲林，主任研究员、博士生导师
收稿日期：2010-03-16 **E-mail：**lisb@dsp.ac.cn

60．**桥侧石柱　晨　外景**
月红、小香给春花松开绑绳。
小香（不忍）
春花。
月红（哽咽）
春姐。
春花抱住月红放声痛哭。

图 1　原始剧本场景描述文本

图 1 中的示例场景文本转换后的内部表示如图 2 所示。该内部表示是本文定义的一种采用上下文无关文法(CFG)进行定义的、具有特定语法格式的剧本场景描述语言。

60．桥侧石柱，晨，外景
（月红、小香给春花松开绑绳。）
小香：（不忍）春花。
月红：（哽咽）春姐。（春花抱住月红放声痛哭。）

图 2　转换后的剧本场景内部表示文本

一个剧本一般由多个场景构成，为让用户可以根据场景的不同方面内容进行视频检索，本文把剧本蕴涵的结构化信息表示为一棵树，如图 3 所示。其中，Script 节点表示剧本，它包含多个场景(Scene 节点)。每个场景又包含发生地点(Location 节点)、发生时间(Time 节点)、场景类型(Type 节点)、场景描述(Description 节点)、场景对应的视频片段(URI 节点)和对白(Dialogues 节点)6 个方面的信息。每个对白节点可包含一句或多句对白(Dialogue 节点)。每句对白包含以下信息：说话角色(Role 节点)，说话时注释(Annotation 节点)，对白内容(Content 节点)，话后说明(Show 节点)，该句对白在视频中的出现时刻(Appear 节点)和该句对白在视频中的消失时刻(Disappear 节点)。本文通过对剧本的内部表示文本进行语法制导翻译，生成如图 3 所示的结构化剧本场景信息。经过上述处理，图 3 中剧本场景结构化信息域的内容除 URI、Appear 和 Disappear 外已全部生成，这 3 个信息域内容的生成方法将在下文讨论。

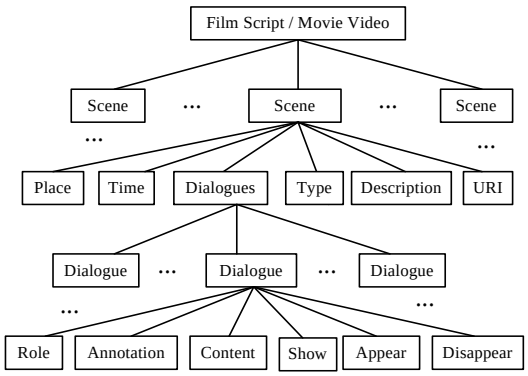


图 3　剧本信息的结构化表示

3　剧本与字幕的匹配方法

假设电影 V 包含的字幕为集合 $T=\{D_1,D_2,\cdots,D_i,\cdots,D_m\}$，集合中的元素 $D_i(1\leqslant i\leqslant m)$ 为一个三元组 $\{DB,DE,DC\}$，DB 为字幕在视频中的出现时刻，DE 为字幕在视频中的消失时刻，DC 是字幕文本。V 对应的剧本用集合 $P=\{S_1,S_2,\cdots,S_i,\cdots,S_n\}$ 表示，集合中的元素 S_i 表示剧本中的第 $i(1\leqslant i\leqslant n)$ 个场景，S_i 可表示为六元组 $\{SI,SL,ST,SP,SS,SD\}$，SI 表示场景的编号，SL 表示场景发生的地点，ST 表示场景发生的时段(晨、日或夜等)，SP 为场景类型(内/外景)，SS 为场景的描述文本，SD 为场景中的对白。SD 是一个集合 $\{d_1,d_2,\cdots,d_i,\cdots,d_p\}$，其中，$d_i$ 为一个五元组 $\{SDR,SDA,SDC,SDP,SDD\}$，SDR 为说话角色名，SDA 是说话者说出该对白时的一些注释信息，SDC 是对白的文字内容，SDP 为该条对白在字幕中的出现时刻，SDD 为该条对白在字幕中的消失时刻。设 V 中所有字幕文字按其出现顺序构成集合 $X=\{DC_1,DC_2,\cdots,DC_i,\cdots,DC_m\}$，为便于书写，改写为 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_m\}$ 并以 $x_i\cup x_{i+1}$ 表示第 i 条和第 $i+1$ 条字幕文字的连接，只有连续字幕文字才能进行连接操作。剧本 P 中所有对白中的文字按其出现顺序构成集合 $Y=\{SDC_{1,1},SDC_{1,2},\cdots,SDC_{2,1},SDC_{2,2},\cdots,SDC_{i,j},\cdots,SDC_{n,p}\}$，其中，$SDC_{i,j}$ 表示第 i 个场景中第 j 个对白的文字内容，为便于书写，将集合改写为 $Y=\{y_{1,1},y_{1,2},\cdots,y_{2,1},y_{2,2},\cdots,y_{i,j},\cdots,y_{n,p}\}$。$y_{i,j}$ 可能被拆成多条连续字幕，为获取对白的出现和消失时刻，必须找出满足式(1)的所有 k 和 w，根据这 2 个值即可获得 $y_{i,j}$ 所在对白在视频中的出现时刻 $SDP_{i,j}=DB_k$ 和消失时刻 $SDD_{i,j}=DB_w$。获得剧本中对白的出现和消失时刻后，假设场景 $S_i(1\leqslant i\leqslant n)$ 中有对白，其中，第 1 个对白的文字为 $y_{i,1}$ 且 $y_{i,1}=x_k\cup x_{k+1}\cup\cdots\cup x_l$，最后一个对白的文字为 $y_{i,p}$ 且 $y_{i,p}=x_{l+h}\cup x_{l+h+1}\cup\cdots\cup x_u$，则场景 S_i 对应的视频片段的初始边界为 $[DB_k,DE_u]$（DB_k 开始时刻，DE_u 为结束时刻）。对于正式出版的剧本，场景中包含的对白与字幕基本上是一一对应的(字幕和对白中的文字在个别字词上存在差异，使用字符串近似匹配判定 2 个串是否匹配)，此时使用上述方法可以为每个包含字幕场景所对应视频片段确定初始分割边界。但对于某些非正式出版的剧本，可能并不列出影片中出现的所有对白，此时如果场景 S_i 的第 1 个对白和最后一个对白存在，则根据上述方法仍然能确定其对应视频片段的初始分割边界，否则场景片段的分割将更多地依赖视频内容分析。

$$\forall y_{i,j}\ (\exists_{1\leqslant k\leqslant l\leqslant m}(x_k\cup x_{k+1}\cup x_{k+2}\cup\cdots\cup x_l=y_{i,j}))\quad(1)$$
$$1\leqslant i\leqslant n,1\leqslant j\leqslant p$$

4　视频场景片段抽取与索引

经过上文处理已经为剧本中包含对白的场景生成了对应视频片段的初始分割边界，但由于某些场景可能不包含对白且场景之间可能存在场景切换对应的视频片段，因此为比较准确地分割场景视频片段，本文利用基于视频内容分析的方法进一步确定剧本中所有场景的分割边界。上文已经假设场景 S_i 的初始分割边界为 $[DB_k,DE_u]$，另设 S_i 后第 1 个包含对白的场景为 $S_{i'}$，所对应视频片段的初始分割边界为 $[DB_{k'},DE_{u'}]$，用 $N=|S_{i'}-S_i|$ 表示 S_i 到 $S_{i'}$ 之间无对白场景的个数，则视频段 $[DE_u,DB_{k'}]$ 要确定的分割点数量为 $N+1$，设 $b_1,b_2,\cdots,b_{N+1}$ 为基于视频内容分析找到的场景切换点，则场景 S_i 对应视频片段的结束时刻为 b_1，场景 $S_{i'}$ 对应视频片段的开始时刻为 b_{N+1}，中间每个无对白场景对应的视频片段的分割边界为 $b_i,b_{i+1}](1\leqslant i\leqslant N)$。为找到上述 $N+1$ 个场景切换点，本文统计了 $[DE_u,DB_{k'}]$ 之间的视频帧的帧间差异，并选出差异最大的 $N+1$ 个作为切换点。颜色不易受对象运动的干扰，是一个在视频内容分析中广泛使用的视觉特征，为大部分的视频场景片段抽取算法所采用，因此，本文采用比较颜色直方图的方法确定帧间差异，使用式(2)计算 2 个连续视频帧 n 和 $n+1$ 之间的颜色分布差异情况。

$$D=\sum_{p=1-L}^{L-1}\left|C_n^1(p)-C_{n+1}^1(p)\right|+\sum_{p=2(1-L)}^{2(L-1)}\left|C_n^2(p)-C_{n+1}^2(p)\right|+\sum_{p=0}^{3(L-1)}\left|C_n^3(p)-C_{n+1}^3(p)\right| \quad (2)$$

式(2)假定原 RGB 颜色分量的量化等级为 L，$C_n^i(p)(1\leqslant i\leqslant 3)$ 表示变换后颜色空间中第 i 个颜色分量量化等级为 p 的像素点个数。

使用上述方法确定所有场景的边界后，可从视频文件分割出相应的视频片段，将该视频片段的存取地址作为图 3 中场景 URI 域的值。至此，剧本(即视频)的树形结构化表示中各个域的值生成完成，基于图 3 的视频结构化表示文档可以作为视频检索的索引。

5 实验结果

本节对上文介绍的视频场景分割方法进行性能分析。在电影《芙蓉镇》和《牧马人》中选取包含前 15 个场景的视频片段作为测试视频，下文称这 2 个视频片段为 test1 和 test2，这 2 部电影的剧本可从公开出版的《谢晋电影选集》中获得。本文采用查全率和查准率衡量场景分割的准确性。测试用机器 CPU 为 Intel Pentium Ⅲ 866 MHz，其内存 512 MB。

对于查全率，由于剧本提供了明确的所要分割的视频场景片段数目信息，因此在剧本准确的情况下，本文方法可以保证查全率为 100%，不再讨论。对于查准率，本文方法主要取决于剧本中存在对白的场景所包含对白的完整性。图 4 给出了 test1 和 test2 在连续且不包含对白的场景数量发生变化时，查准率的变化情况。

本文通过去掉剧本中某些场景包含的对白获得 test1 和 test2 对应的包含多个连续且无对白场景的剧本，test1 对应的原始剧本中包含 15 个场景,其中有 3 个不连续的无对白场景。test2 对应的原始剧本中包含 15 个场景，其中包含 2 个连续无对白场景。从图 4 可以看出，在剧本未经改动时，查准率达到 100%，但在连续无对白场景的数量增加时，查准率随之下降，其原因是对于连续多个不包含对白的场景，其场景边界完全由基于帧间颜色直方图差异的方法确定，而该方法在确定少量场景边界中的切换点时是可行的，并不能用于视频场景分割。在连续不包含对白的场景数量不超过 2 的情况下，从图 4 可知本文方法具有 70%以上的查准率，经过对大量剧本的统计发现，在剧本中一般不会出现连续 2 个以上的无对白场景，这使本文方法可以保持较高的查准率。

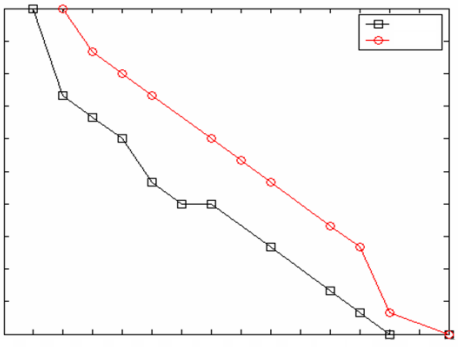


图 4 场景分割准确率变化情况

图 5 给出了在连续无对白场景的数量增加时，视频分割所消耗时间的变化情况。可以看出，随着连续无对白场景数量的增加，视频分割所消耗的时间逐渐递增，在连续无对白场景数量为 15 时，本文方法必须分析整个视频片段所包含的所有视频帧，消耗的时间最多。此时进行的是基于颜色直方图的镜头分割。完全依赖视频内容分析的场景分割方法一般先分割出镜头，然后通过对镜头进行聚类获得场景，镜头分割是其前期步骤。因此，假设镜头聚类消耗的时间很少，可以忽略，但此类方法消耗的时间仍比本文方法多，从图 5 可以看出，在剧本场景对白未删减时，其消耗的时间为本文方法的 2 倍以上。

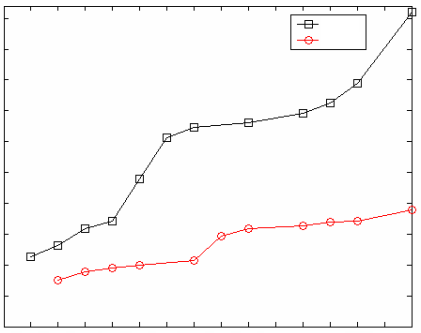


图 5 场景分割所消耗时间变化情况

6 结束语

本文方法在视频场景分割的精度和速度方面具有较高性能，可以自动为分割的视频场景片段生成文本，具有较大应用价值。

参考文献

[1] Zhai Yun, Shah M. A General Framework for Temporal Video Scene Segmentation[C]//Proc. of International Conference on Computer Vision. Beijing, China: [s. n.], 2005: 1111-1116.

[2] Tavanapog W, Zhou Junyu. Shot Clustering Techniques for Story Browsing[J]. IEEE Transactions on Multimedia, 2004, 6(4): 517-526.

[3] 张振原, 路 红. 一种基于视频结构的场景分割方法[J]. 中国图象图形学报, 2007, 12(10): 1913-1916.

[4] Chen Lianghua, Lai Yuchun, Liao H Y M. Movie Scene Segmentation Using Background Information[J]. Pattern Recognition, 2008, 41(3): 1056-1065.

[5] Merlino A, Morey D, Maybury M T. Broadcast News Navigation Using Story Segmentation[C]//Proc. of ACM International Conference on Multimedia. Seattle, Washington, USA: [s. n.], 1997: 381-390.

[6] Xie Lexing, Xu Peng, Chang Shih-Fu, et al. Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models[J]. Pattern Recognition Letter, 2004, 25(7): 767-781.

编辑 陈 晖