第 35 卷　第 1 期　　计　算　机　工　程　　2009 年 1 月
Vol.35　No.1　　Computer Engineering　　January 2009
·博士论文·　　文章编号：1000—3428(2009)01—0033—03　　文献标识码：A　　中图分类号：TP391

新闻视频结构化浏览与标注系统

刘安安[1,2]，李锦涛[2]，张勇东[2]，唐　胜[2]，杨兆选[1]，吴佳鹏[1]
(1. 天津大学电子信息工程学院，天津 300072；2. 中科院计算技术研究所智能信息处理重点实验室，北京 100080)

摘　要：阐述一种新颖的新闻视频结构化浏览和标注系统。应用基于时空切片分析的新闻主播检测方法和基于颜色直方图的镜头分割方法实现新闻视频的结构化。通过自动语音识别技术和特定语义概念模型的建立实现了对主播场景的文本信息标注和对新闻故事镜头的语义概念标注。该系统有利于用户根据个人爱好进行新闻视频的浏览和编辑，有效实现新闻视频的索引和浏览。

关键词：新闻；时空切片；语义概念；自动语音识别

Hierarchically Browse and Annotation System for News Video

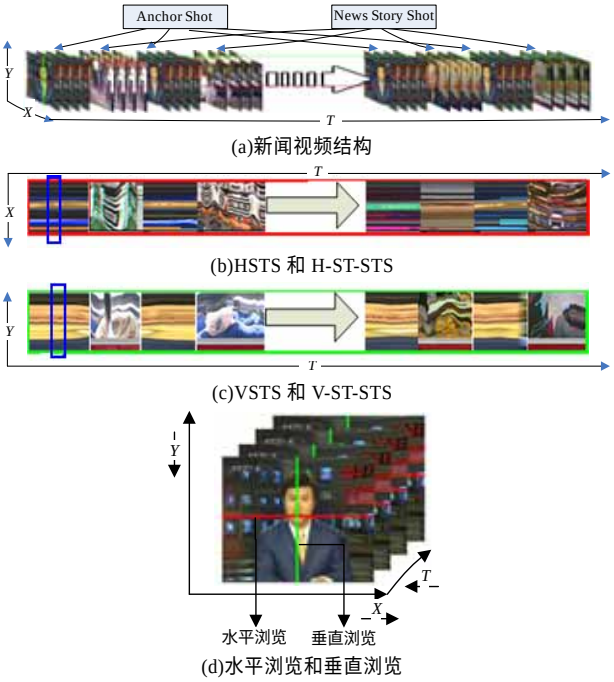
LIU An-an[1,2], LI Jin-tao[2], ZHANG Yong-dong[2], TANG Sheng[2], YANG Zhao-xuan[1], WU Jia-peng[1]
(1. School of Electronic and Information Engineering, Tianjin University, Tianjin 300072;
2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

【Abstract】This paper presents an innovative hierarchically browsing and annotating system for news video. With the anchorperson detection algorithm based on Spatio-Temporal Slice(STS) analysis and the shot boundary detection method based on color histogram, structuralization on video is implemented. With Automatic Speech Recognition(ASR) technique and semantic concepts models, the anchorperson scenes with text information are respectively annotated and the news story scenes with visual semantic concepts are labeled. Therefore, the system can realize the interest-oriented browse for viewers and facilitate video edit for editors by providing effective indexing and browsing method for news video.

【Key words】news; Spatio-Temporal Slice(STS); semantic concept; Automatic Speech Recognition(ASR)

1　概述

随着视频数据的海量增长，人们迫切需要先进的视频处理技术从而实现有效的视频索引、浏览以及检索。为了解决这个难题，从 20 世纪 80 年代起，大量的研究人员开始从事基于内容的多媒体内容分析。在该研究领域中，由于新闻视频具有独特的结构特点(如图 1 所示)而成为重要的研究对象。



(a)新闻视频结构
(b)HSTS 和 H-ST-STS
(c)VSTS 和 V-ST-STS
(d)水平浏览和垂直浏览

图 1　新闻视频结构及相应的时空切片

在文献[1]中，研究人员通过结合视频和音频等多模态信息将新闻视频分割成为场景、事件、镜头等不同粒度的视频结构单元，以方便视频内容的管理。此外，为了挖掘视频的语义信息，文献[2]根据新闻视频特点，从台标识别、栏目识别、字幕识别、播音识别等方面对视频语义的提取进行了探索。

尽管在相关研究中已经取得了很大进展，但是在新闻视频结构和语义分析中仍存在如下问题：

(1)目前的多模态融合的视频结构化方法过于复杂且准确率较低，普遍的准确率在 90%左右。

(2)对于新闻视频语义的分析仍停留在低级和中级语义，并不能支持有效的语义标注和检索。

为了解决上述问题，笔者开发了一种新颖的新闻视频结构化浏览和标注系统。该系统的创新性在于：

(1)根据新闻视频的结构特征，提出了一种准确性高，鲁棒性强的基于时空切片模式分析的新闻主播检测方法，从而实现了将新闻视频分割为主播场景和新闻故事场景。

(2)对于视觉内容变化不大的主播场景，通过自动语音识别技术实现对该类场景基于文本的标注；对于视觉内容变化

基金项目：国家“973”计划基金资助项目(2007CB311100)；国家“863”计划基金资助项目(2007AA01Z416)；天津市自然科学基金资助重点项目(07JCZDJC05800)

作者简介：刘安安(1982－)，男，博士研究生，主研方向：多媒体检索；李锦涛，研究员、博士；张勇东，副研究员、博士；唐　胜，助理研究员、博士；杨兆选，教授、博士；吴佳鹏，博士研究生

收稿日期：2008-06-15　**E-mail：**liuanan@ict.ac.cn

明显的新闻故事场景，根据基于视觉低层特征建立的语义概念模型实现对该类场景基于语义概念的标注。

2 新闻视频结构化浏览和标注系统

笔者开发了一种新颖的新闻视频结构化浏览和标注系统。该系统由视频结构化模块、文本和语义标注模块以及交互界面 3 部分组成。本部分中仅对交互界面进行介绍。

如图 2 所示，该系统界面包括 5 个部分：概念标识子窗口 I(左上)，文件选择子窗口 II(左中)，文件播放子窗口 III(左下)，结构化浏览和语义概念标注子窗口 IV(右上)，新闻内容摘要子窗口 V(右下)。

图 2 新闻视频结构化浏览和标注系统交互界面

为了直观地浏览新闻视频的结构，在子窗口 IV 中将新闻视频通过二维笛卡尔坐标系展开。其中，垂直方向为一个新闻故事每个镜头的第 1 个关键帧；水平方向表示按时间顺序排列的各个主播场景和新闻故事场景。为了标注各个新闻故事镜头所包含的视觉语义概念，在子窗口 I 中将不同的颜色对应于不同的语义概念，在各个关键帧下通过颜色条带来表示该镜头所包含的语义概念。同时，用户也可以将鼠标指向各个关键帧，此时会显示该镜头所包含的语义。对于各个主播场景，可以用鼠标单击关键帧，此时在子窗口 V 中将显示该新闻部分的内容摘要，即文本标注。通过该交互界面，用户可以根据新闻视频结构浏览、视觉语义概念标注信息和新闻内容摘要信息选择并点击感兴趣的镜头，此时将在子窗口 III 中从所选镜头开始播放。因此，该系统为用户提供了有效的新闻视频索引和浏览工具。

3 新闻视频结构化

新闻视频通常由主播场景和故事场景组成。由于一段视频中重复出现的主播段通常具有很强的视觉相似性，而其他镜头仅与同一故事单元中临近的镜头具有相似性，因此新闻主播场景类比其他类别包含更多的成员。

根据新闻视频自身特点，可以通过主播场景的检测实现新闻视频的场景分割，从而既避免了目前自底向上的场景分割算法(如文献[3]中所述的图分割算法)的复杂性，又提高了检测的准确率。

在本系统中，通过基于时空切片模式分析的主播检测算法实现了新闻视频的场景分割；在此基础上，对于各新闻故事场景，通过颜色直方图比较的方法进行镜头分割和关键帧提取，从而实现了新闻视频的结构化。

3.1 新闻主播场景检测

3.1.1 定义及规则

为了便于对本算法的介绍和理解，本节阐述一些重要的概念。

(1)定义

1)时空切片。

时空切片是由视频图像序列的连续帧中相同位置的像素条带按时序组成的一幅图像。由于像素条带选择方式的不同，时空切片具有多样性。

2)短时时空切片。

一个时空切片可以平均分成一组时间上连续且等宽的时空切片片段，每个时空切片片段被称为短时时空切片。本算法采用了由水平像素条带组成的水平时空切片和由垂直像素条带组成的垂直时空切片，以及对应的短时时空切片，如图 1 所示。

3)新闻主播场景。

新闻主播场景通常包含演播室背景和一或两个新闻主播，同时可能存在摄像机角度和台标的变化。

4)新闻主播场景组。

新闻主播场景组由一个视频中所有新闻主播场景组成。

(2)规则

基于上述对新闻视频特点的分析，根据如下规则检测新闻主播场景：在一个完整新闻视频中，新闻主播场景组是包含视觉内容相似的视频结构单元最多的组，他的每个成员对应一个新闻主播场景。因此，从理论上讲，对应于新闻主播场景的短时时空切片可以聚为一类，并且这个类应该包含最多的元素。

3.1.2 新闻主播检测

本节将介绍基于时空切片的新闻主播检测算法。该算法包含如下 3 个部分：时空切片模式分析和特征提取，聚类，信息融合。

(1)时空切片模式分析和特征提取

通过仔细观察图 1(b)和图 1 (c)中所示时空切片可知，时空切片颜色和纹理的不连续性意味着新的镜头的出现。通过对前面所阐述的新闻主播镜头的特性的分析可见，主播出现的视频帧通常具有很强的视觉相似性。因此，从主播镜头中提取的短时时空切片通常变化很细微，可以将相似的短时时空切片聚为一类。

显而易见，颜色特征和纹理特征很好地表征了时空切片的特性。对于每个短时时空切片，首先将其等分为 4×4 子块并提取各个子块的 32 维颜色直方图来表征局部颜色特征。然后，计算颜色的一阶矩、二阶矩和三阶矩来表征全局颜色特征。此外，计算文献[4]中介绍的边缘直方图描述子来表征纹理特征。因此，每个短时时空切片由 515 维颜色特征和 150 维纹理特征组成的高维视觉特征向量表征。

(2)短时时空切片聚类

K 均值算法是一种简单而有效的聚类算法。笔者利用该方法对由视觉特征向量表征的水平和垂直短时时空切片分别进行聚类。

聚类后，分别将同类中时间连续的水平和垂直短时时空切片合并作为类中的一个成员。这样，每个成员对应一个镜头(Shot)。因此，每个类的组成可以用下式表示：

$$\begin{aligned} Cluster_i^H &= < Shot_{i1}^H, Shot_{i2}^H, \cdots, Shot_{iR}^H > \\ Cluster_j^V &= < Shot_{j1}^V, Shot_{j2}^V, \cdots, Shot_{jS}^V > \end{aligned} \quad (1)$$

其中，$Cluster_i^H$ 和 $Cluster_j^V$ 分别表示水平切片聚类结果中的第 i 个类和垂直切片聚类结果中的第 j 个类；R 和 S 分别表示聚类结果中的元素个数。

(3)信息融合

为了综合利用水平和垂直信息使检测更准确，融合水平和垂直切片的聚类结果来作最终判决。根据 2.1.1 节中所阐述的检测规则，仅对水平和垂直方向包含最多元素的类(即 $Cluster_{\max}^H$ 和 $Cluster_{\max}^V$)进行融合。$Cluster_{\max}^H$ 中的 $Shot_i^H$ 和 $Cluster_{\max}^V$ 中的 $Shot_j^V$ 的相似度计算如下：

$$Sim<Shot_i^H,Shot_j^V>=\frac{\min(T_{\text{end}}^H,T_{\text{end}}^V)-\max(T_{\text{start}}^H,T_{\text{start}}^V)}{\max(T_{\text{end}}^H,T_{\text{end}}^V)-\min(T_{\text{start}}^H,T_{\text{start}}^V)} \quad (2)$$

其中，$T_{\text{start}}^H,T_{\text{end}}^H,T_{\text{start}}^V,T_{\text{end}}^V$ 分别表示 $Shot_i^H$ 和 $Shot_j^V$ 的起始时间和终止时间；min和max分别表示取最小值和最大值操作。如果 2 个镜头的相似度大于阈值Th_1(经验值，本实验中设定为 0.5)，则这 2 个镜头称为“对应”镜头；否则为“非对应”镜头。如果相似度值小于 0，则将其修正为 0。于是，$Cluster_{\max}^H$ 和 $Cluster_{\max}^V$ 的相似度通过下式计算：

$$Sim<Cluster_{\max}^H,Cluster_{\max}^V>=\frac{1}{\min(R,S)}\sum_{i=1}^{R}\sum_{j=1}^{S}Sim<Shot_i^H,Shot_j^V> \quad (3)$$

融合方法如下：

(1)如果 $Cluster_{\max}^H$ 和 $Cluster_{\max}^V$ 的相似度大于阈值Th_2(经验值，本实验中设定为 0.5)，则将它们融合为一类。此时，如果一个镜头在另一个类中不包含对应的元素，则将其作为融合得到的新的类中的一个成员；否则，将对应镜头的最早开始时间和最晚结束时间中间部分对应的镜头作为新类的一个成员。于是，得到的新类即为新闻主播场景组。

(2)如果 $Cluster_{\max}^H$ 和 $Cluster_{\max}^V$ 的相似度不大于阈值Th_2，则不将 2 类进行融合。此时，由于视频不包含明显的主播镜头，无需进行主播场景检测。

3.1.3 实验结果

为了证明本文所述方法对于不同的视频源和先进的视频编辑技术的鲁棒性，从 TRECVID2005 和 TRECVID2006 (TRECVID 是由美国 NIST 组织支持的信息检索领域的顶级评测)的测试集中选择了 40 个新闻视频用于测试。有关测试集的详细信息可以参照表 1。表 2 中所示实验结果证明了该新闻主播检测算法具有很高的准确性和鲁棒性。

表 1 测试数据

视频源	数量	长度/min
NTDTV (20041101_120001~20041109_120100)	10	约 30
CCTV (20041105_150000~ 20041114_150000, 20051201_145800~ 20051217_145800)	30	约 10

表 2 实验结果

视频源	标注的主播场景数	检测的主播场景数	错检	漏检	查准率/(%)	查全率/(%)
NTDTV	199	199	0	0	100	100.00
CCTV	502	500	0	2	100	99.60

3.2 镜头分割和关键帧提取

由于新闻故事场景中的镜头多为切变，因此采用了颜色直方图比较的镜头分割方法和无监督聚类的关键帧提取方法。具体方法可以参考文献[5]。

4 新闻视频的标注

由于主播场景是对新闻故事的简介，因此通过自动语音识别技术对其进行文本标注；而对于视觉内容丰富的新闻故事场景，利用基于视觉低层特征建立的语义概念模型对其进行语义概念标注。

4.1 基于文本的新闻主播场景标注

笔者使用了微软亚洲研究院开发的普通话语音工具箱进行自动语音识别。首先从视频文件中提取音频流，并提取音频头文件中的采样率、编码比特数、通道数等参数对语音识别组建进行初始化，然后调用识别引擎进行识别。

在测试中，在 TRECVID 评测的 Highlevel Feature Extraction 2006 的数据中随机选取了 5 个中文视频(包括 CCTV 和 NTDTV 2 个中文新闻台)和 5 个英文视频(包括 CNN 和 NBC 2 个英文新闻台)，并将这些新闻视频中的主播镜头人工提取作为测试数据。然后通过上述方法进行自动语音识别。实验结果如下：中文 ASR 准确率为 75%；英文 ASR 准确率为 60%。

4.2 基于语义概念模型的新闻故事场景标注

新闻故事场景通常包含如下 3 方面内容：采访场景，人物活动场景，自然风光场景。根据新闻视频所包含的语义概念及 2007 年视频检索领域顶级评测TRECVID中的高级语义检索项目对视频语义概念的分类，选取了人脸(face)、采访(interview)、人(person)、人群(crowd)、室外(outdoor)、建筑(building)共 6 个视觉语义概念对新闻故事场景进行标注。其中，对于前 3 个易于通过低层特征表征的概念分别建立了特定的模型来提高检测结果；而对于后 3 个概念，利用通用概念模型训练的架构进行模型建立。具体方法可以参考笔者在 2006 年TRECVID评测的子项目RUSHES的研究报告[5]。

5 结束语

本文阐述了一种新颖的新闻视频结构化浏览和标注系统。通过视频的结构化和语义标注，建立了交互式的新闻视频浏览系统。该系统的应用为用户提供了有效的新闻视频索引和浏览工具。

参考文献

[1] 史迎春, 方鹏飞, 周献中. 综合利用声视特征的新闻视频结构化模型[J]. 计算机工程与应用, 2004, 40(32): 99-101.

[2] 史迎春, 王 韬, 周献中. 基于语义的新闻视频检索研究[J]. 计算机工程, 2004, 30(16): 155-157.

[3] Rasheed Z, Shah M. Detection and Representation of Scenes in Videos[J]. IEEE Trans. on Circuits and Systems for Video Technology, 2005, 7(6): 1097-1105.

[4] Park D K, Jeon Y S, Won C S, et al. Efficient Use of Local Edge Histogram Descriptor[C]//Proc. of the ACM Workshops on Multimedia. Los Angeles, CA, USA: [s. n.], 2000.

[5] Tang Sheng, Zhang Yongdong, Li Jintao, et al. TRECVID 2006 Rushes Exploitation by CAS MCG[C]//Proc. of TRECVID Workshop. Gaithesburg, USA: [s. n.], 2006.