文章编号：2095-2163(2024)01-0070-07 中图分类号：TP399 文献标志码：A

# 基于多模态相似融合的新闻视频故事分割算法

吴 培，周激流
（四川大学 电子信息学院，成都 610065）

**摘 要**：新闻视频数量的不断增加，为准确分割用户感兴趣的新闻视频，本文提出了一种基于多模态相似融合的新闻视频故事分割算法。首先，通过选定视频切割点获取候选新闻故事单元边界，将视频分成音频流和视频流；其次，选择静音区间为音频候选切分点，主持人镜头帧和主题字幕帧作为视频候选切分点，根据候选切分点获得新闻故事基本单元，利用语义相似性分析各单元内容进行合并或独立分离，得到最终新闻故事；最后，采用人脸识别、YOLOv5 来进行主题字幕检测、语义相似性合并或独立新闻故事基本单元，使得新闻故事边界划分更为准确。该新闻视频故事分割算法在《新闻联播》视频中查全率和查准率分别达到了 97.17% 和 98.19%，为新闻视频导航、检索等应用提供辅助准备。

**关键词**：新闻故事基本单元；主题字幕；人脸识别；YOLOv5；语义相似性

# News video story segmentation algorithm based on multi-mode similarity fusion

WU Pei, ZHOU Jiliu
(College of Electronic Information, Sichuan University, Chengdu 610065, China)

**Abstract**: With the increasing number of news videos, in order to accurately segment news videos of interest to users, this paper proposes a news video story segmentation algorithm based on multimodal similarity fusion. Firstly, by selecting video cutting points to obtain candidate news story unit boundaries, the video is divided into audio and video streams; Secondly, select the silent interval as the audio candidate segmentation point, and the host lens frame and theme subtitle frame as the video candidate segmentation points. Based on the candidate segmentation points, obtain the basic units of the news story, and use semantic similarity analysis to merge or separate the content of each unit separately to obtain the final news story; Finally, facial recognition and YOLOv5 are used for topic subtitle detection, semantic similarity merging, or independent news story basic units to make news story boundary division more accurate. The recall and precision of the news video story segmentation algorithm in CCTV News video reached 97.17% and 98.19% respectively, providing auxiliary preparation for news video navigation, retrieval and other applications.

**Key words**: basic unit of news story; topic subtitles; Face recognition; YOLOv5; semantic similarity

## 0 引 言

近年来，随着网络技术的不断发展和计算机、手机、平板等硬件性能的不断提高，人们对视频的需求呈现出快速增长的趋势。视频信息因其多样性和复杂性而备受关注，同时也因其多媒体特性而被广泛应用于日常生活的各个领域。由于视频内容的数量和种类不断增加，用户往往无法在有限的时间内浏览所有的视频，在浏览视频时需要花费大量时间来寻找自己感兴趣的内容。在视频创作过程中，对视频进行剪辑以分割出用户感兴趣的内容变得至关重要。传统的人工视频分割方法已经不能满足快速产出视频的需求，视频智能化、自动化快速剪辑成为主流趋势。然而视频内容结构复杂多样，采用一种通用的方法对所有的视频进行剪辑分段是不现实的，因此本文针对特定的视频类别，以提高剪辑效率和准确性为目标。

新闻视频是们获取信息的重要途径之一，在海量信息中迅速、准确地定位所需内容已成为当今亟待解决的课题。相对于其他类型的视频数据，新闻视频呈现出明显的结构特征，其内部包含多个独立的新闻故事单元。所谓新闻故事单元是在一段新闻视频节目中，时间和空间上呈现出相对连贯性的一系列视频镜头，构成一个完整的故事单元[1]。在新

**作者简介**：吴 培(1998-)，男，硕士研究生，主要研究方向：通信与信息系统。
**通讯作者**：周激流(1963-)，男，博士，教授，博士生导师，主要研究方向：图像处理、计算智能、分数阶微积分理论及其在信号处理中的应用等。Email：zhoujl@ scu.edu.cn

**收稿日期**：2023-08-28

闻故事单元发生转换时，通常意味着内容发生较大变化。以《新闻联播》为例，不同领域如农业和体育的新闻包含了截然不同的场景，多种人物的特写镜头，不同主题的字幕展示以及相关采访的片段，这些独立的片段可以被视为构成新闻报道中不同新闻故事单元的组成部分。

本文针对如何充分提取并综合各方面信息来分析新闻视频的问题，提出了一种基于多模态相似融合的新闻视频故事分割算法，自动的从长视频中分割出独立的新闻视频故事单元，为后续的短视频处理做准备工作；多特点融合选取视频候选切分点，利用主持人特征、主题字幕、语义特征和音频特征等多种特征提升新闻故事分割的准确率；在时域上借助语义相似性分析各新闻故事基本单元内容的相似性，对基本单元进行合并或独立分离。该新闻视频故事分割算法在《新闻联播》视频中，较现有模型获得了更好的性能。

## 1　视频结构特征

视频通常被定义为一系列连续的图像帧，以一定的速率连续播放，从而在人眼中产生动态视觉效果。根据视频的结构特点，将视频数据划分为 4 个层次结构：故事层、场景层、镜头层和图像帧层，如图 1 所示[2]。

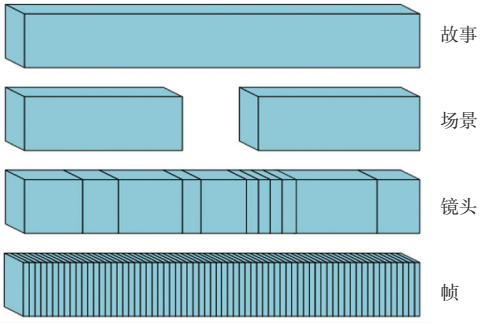


**图 1　视频结构分层图**

**Fig. 1　Video structure hierarchy diagram**

图像帧是从视频流中捕获的单个图像单元，即视频的最微小成分。通过在特定帧位置中断，可以单独获取这个图像单元。镜头则是一系列持续捕捉的画面，通常是数以百计至千计的镜头在一个视频段中交替呈现，彼此通过剪辑或过渡手法相互转换。场景则常由多个时间上连续且语义相关的镜头构成，呈现某一特定事件或特定环境的内容[3]。故事则由多个场景组合而成，从内容角度完整地叙述与其主题相关的事务。多个语义独立的故事构成视频数据，通过研究视频结构特点，能够精确定位视频切分点，实现对不同主题视频的语义层面分析。

新闻视频由多个新闻故事单元组成。新闻故事单元是指在新闻内容上相关，描述一个完整事件的视频片段[4]。研究发现，新闻故事单元主要呈现出 3 类不同的形式：第一类只包含主持人镜头；第二类只包含内容镜头；第三类既包含主持人镜头又包含内容镜头，分别表述一件完整的新闻故事信息[5]。新闻视频结构特征如图 2 所示，蓝色块为主持人镜头；绿色块为内容镜头；纵向的粗实线段间为 1 个新闻故事。通过观察可以发现：当内容镜头结束，主持人镜头开始时，这两者之间一定存在新闻故事边界，可作为新闻视频切分的依据，得到新闻基本处理单元。

在新闻的基本处理单元中，往往包含有多个独立的新闻故事单元。对于大量新闻视频的观察表明，每个独立的新闻故事单元通常仅会包含一个主题字幕，而这个主题字幕能够简要概括整个新闻故事单元所涵盖的内容。此外，在主持人从一个新闻故事单元切换至另一个新闻故事单元的过程中，往往伴随着明显的播报声音停顿。这两种特征可以用来进一步处理新闻基本单元，更加准确的完成对新闻视频的分割。

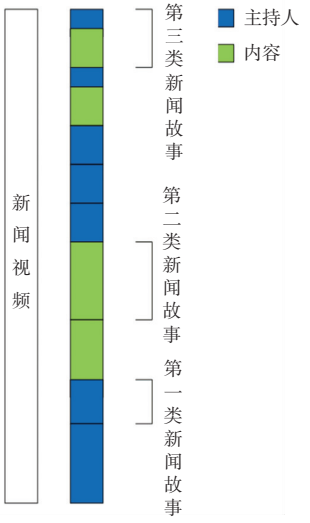


**图 2　新闻视频结构特征**

**Fig. 2　Structural characteristics of news video**

## 2　算法模型

### 2.1　算法模型

多模态特征融合的新闻故事单元分割算法模型如图 3 所示，在新闻视频部分，首先对原始视频进行预处理，将其分为视频流和音频流；通过镜头边界检测算法，将视频分成不同的镜头片段，使用主持人检

测算法来识别视频中的主持人,并对其进行跟踪和定位;同时,使用主题字幕检测算法来提取视频中的文本信息,判断是否为主题字幕,通过对音频流中静音帧的检测,筛选出音频候选点;在视频切分点融合分析阶段,根据主持人镜头、主题字幕镜头和音频候选点对新闻进行故事划分;最后,分析各新闻故事基本单元的语义相似性,根据相似性对各新闻故事基本单元进行合并或是分离得到独立的新闻故事单元。

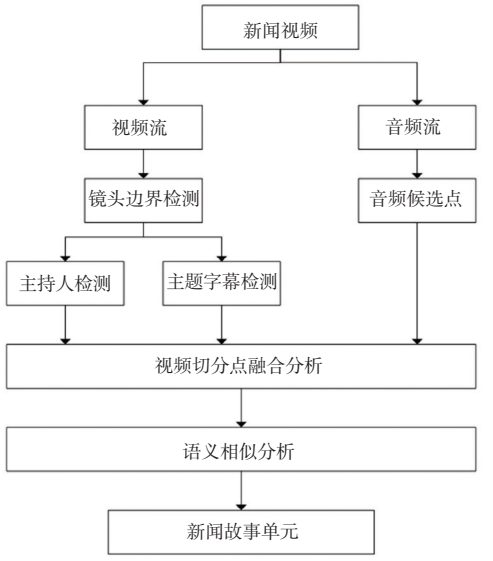


**图 3 算法模型**

**Fig. 3 Algorithm model**

## 2.2 新闻视频主持人镜头检测

新闻视频中,主持人镜头的开始标志着一个新闻故事单元的结束,同时也是一个新的新闻故事的开始,因此主持人镜头是切分新闻故事单元的一个重要依据[6]。在新闻联播中,主持人的名单是固定的,可以依据人脸识别检测主持人镜头帧。

人脸识别是根据人脸图像进行身份识别的一项生物特征识别技术[7]。深度学习由多层自编码神经网络预训练,然后通过采集人脸图片或视频等面部数据进一步优化神经网络权值的深度置信网络(DBN)。DeepFace 是 Python 轻量级人脸识别和人脸属性分析框架,其采用了一种基于检测点的人脸检测方法。在人脸检测中,先选择 6 个基准点,2 只眼睛中心点、1 个鼻子点和 3 个嘴上的点,局部二值模式(LBP)特征用支持向量机(SVR),获取标记点。本文将 DeepFace 模型用于人脸验证,识别主持人镜头,将获取到的主持人镜头帧出现的时间点标记下来,比较两个相邻时间点的差值,若差值非常小则忽略,若差值较大,说明两相邻时间点出现了内容镜头,将后一个时间点作为切分新闻视频的依据,由此得到新闻基本处理单元,获取新闻基本处理单元算法框架如图 4 所示。

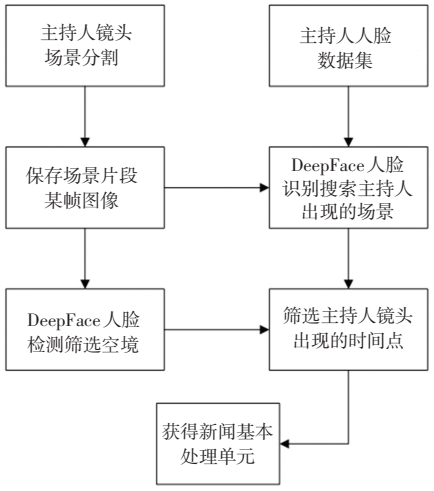


**图 4 获取新闻基本处理单元算法框架图**

**Fig. 4 Basic processing unit algorithm framework of news acquisition**

## 2.3 新闻视频主题字幕帧检测

通过对大量的新闻视频观察,可发现新闻视频中字幕文本主要有主题字幕帧、会话字幕帧和其他字幕帧 3 类[5]。字幕文本分类如图 5 所示。


**图 5 字幕文本分类**

**Fig. 5 Subtitle text classification**

在新闻视频中,主题字幕不仅和新闻基本故事单元存在一一对应的关系,还是对新闻基本故事单元内容的高度概括。在《新闻联播》当中,主题字幕通常出现在一段新的新闻基本故事单元的开始,持续几秒到几十秒,是识别新的新闻基本故事单元的重要标志,因此,主题字幕帧的检测对新闻视频故事分割有着重要的作用[8]。

### 2.3.1 视频帧预处理

新闻视频帧率较高,视频帧前后变化不大,为了降低视频处理的复杂度和计算量,将新闻视频流以每隔 25 帧进行采样,转换为图形帧进行保存。

### 2.3.2 YOLOv5 新闻视频主题帧检测

目标检测在计算机领域中可以为图像和视频的

语义理解提供有价值的信息，本质是对所检测的目标进行定位和分析，完成准确高效地找出给定图像中所有感兴趣目标的任务。目标检测框架包括输入端、骨干网络、Neck 网络和输出端。YOLOv5 经过网络结构及训练技巧等方面的改进，检测性能得到巨大提升，其准确、快速、轻量等特点被广泛的应用[9]。目标检测算法数据处理流程：数据收集、数据标注、模型训练、数据识别。

Labelimg 是一款开源的数据标注工具，可以快速、便捷的为主题字幕打上标签，可以先利用 labelimg 制作自己主题字幕数据集。为了缩短网络的训练时间，达到更好的精度，一般再加载预训练权重对神经网络进行训练，本次训练所采用的是预训练权重为 yolov5s.pt，大小为 27 MB。通过修改模型配置文件，训练自己的模型来识别主题字幕。当检测主题字幕模型训练好后，对之前保存的图像帧进行识别，获得主题字幕帧出现的时间。

2.3.3　OCR 主题帧字幕提取

新闻视频主题帧的文本是新闻故事单元的重要信息源，提取其文字信息是划分新闻故事单元的关键。光学字符识别（OCR）通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字，常用于对文本资料的图像文件进行分析处理，获取文字信息。本文采用基于谷歌开源 OCR 引擎 Tesseract 字符识别技术，Tesseract-ocr 引擎主要包括两部分：图片布局分析、字符分割和识别。在图片布局分析中，首先将图像中的图片、表格、文本等内容分开，提取字符识别时的文本区域，删除掉其他信息。将提取的文本区域进行字符切分，Tesseract 使用粗切分和精细切分两个步骤进行字块的切分。切分后的字符区域将传递给识别引擎，输出字符识别结果，其识别效果如图 6 所示。


**图 6　OCR 字符识别**

**Fig. 6　OCR character recognition**

## 2.4　音频分析算法

通过对大量的新闻视频的实验观察，发现在不同的新闻单元之间，主持人的声音会有明显暂停，至少持续 0.3 s[10]。选取短时能量和短时过零率这两个声音信号的物理特征，进行静音分析。通过从音频信号的角度提取出静音帧，精确定位新闻故事单元的边界。

短时能量是帧中采样点的总能量，采用 Hamming 窗对音频进行分帧，每帧 20 ms[11]。设 $x_i(m)$ 为加窗分帧后第 $i$ 帧的音频信号；$E_i$ 为第 $i$ 帧的短时能量，计算如式（1）所示：

$$E_i = \sum_{m=1}^{N} x_i^2(m) \tag{1}$$

其中，$N$ 为第 $i$ 帧包含的音频采样数目。

短时过零率是对信号频率的简单度量，过零率是每秒内信号值过零值的次数。$Z_i$ 为第 $i$ 帧的短时过零率，计算如式（2）所示：

$$Z_i = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \left| sgn(x_i(m+1)) - sgn(x_i(m)) \right| \tag{2}$$

其中，$sgn(\,)$ 为符号函数。

与非静音帧相比，静音帧的短时能量非常小，过零率也很低，因此，利用短时能量特征和短时过零率特征能很好地区分静音帧和非静音帧[5]。$T_e$ 和 $T_z$ 为经验阈值，当 $E_i$ 满足 $E_i \leq T_e$，$Z_i \leqslant T_z$ 时，则该帧为静音帧，否则为非静音帧。

## 2.5　新闻故事单元获取

语义相似性度量两个语篇片段之间意义的等价性，在自然语言处理（NLP）中，测量两个文本片段之间的语义相似性起着重要作用[12]。新闻视频通过主持人镜头检测、主题字幕帧检测、音频分析得到视频镜头的切分点，通过切分点对视频进行切分获得新闻故事基本单元。各个新闻故事基本单元之间部分的语义有很强的相似性，可以合并视为同一个新闻故事单元，而语义相似性较弱的两段新闻故事基本单元可视为相互独立的新闻故事单元。

提取新闻故事基本单元的字幕，若该单元含有主题字幕帧则将主题字幕语义视为该视频单元的语义；若该单元不含主题字幕帧，则提取该视频单元的字幕内容，利用抽取式自动摘要的方法生成该视频单元的语义。HanLP 是一系列模型与算法组成的 NLP 工具包，基于 PyTorch 和 TensorFlow 2.x 双引擎，可对文本进行抽取式自动摘要生成和文本语义相似度分析[13]。通过对相似度值分析，判断是否合并相比较的新闻故事基本单元，最终形成独立的新闻故事单元。

# 3　实验与分析

## 3.1　实验环境

本文所有实验都基于操作系统 Ubuntu20.04，

GPU 为 NVIDIA GeForce GTX2080Ti,为有效验证所提出的新闻故事分割算法,随机选取 2018~2020 年各 15 个 CCTV《新闻联播》视频进行实验。

### 3.2 实验设置

视频共计时长约 23 h,其中共有 683 个新闻故事单元,分辨率为 1 280×720,帧速率为 25 帧/s。主题字幕帧检测实验中,训练和测试都是在 pytorch 开源深度学习框架下进行,从新闻视频中收集了 10 000张字幕图像,分别为主题字幕、会话字幕、其它字幕。数据集采用 PASCAL VOC 格式,并在原数据集的基础上采用翻转和旋转两种方式得到30 000张图片,训练集21 000张,测试集9 000张。训练中的 batch-size 设置为 64,迭代总轮数 epoch 设置为 300,加载预训练权重 yolov5s.yaml。语义相似性分析实验中,使用 HanLP 对主题字幕进行分析,设定语义相似阈值 $T_{similarity}$,当两主题字幕相似度大于 $T_{similarity}$ 则判定两新闻视频故事基本单元属于同一新闻故事单元。通过反复实验,发现 $T_{similarity}$ 设定为0.3时效果最佳。

### 3.3 评价标准

本文采用查全率($R_r$)、查准率($R_p$)和 $F-measure$ 作为新闻视频分割算法性能的评价标准,式(3)~式(5):

$$R_r = N_c/(N_c + N_m) \tag{3}$$

$$R_p = N_c/(N_c + N_f) \tag{4}$$

$$F - measure = 2R_r R_p/(R_r + R_p) \tag{5}$$

其中,$N_c$ 为被正确检出的新闻故事单元数目;$N_m$ 为漏检的新闻故事单元数目;$N_f$ 为被误判的数目。

### 3.4 实验结果

新闻故事分割各个切分点检测的实验结果见表 1。

表 1 详细新闻故事分割实验结果

Table 1 Detailed news story segmentation experiment results

| 类别 | 总数/个 | 正确数/个 | 准确率/% |
|---|---|---|---|
| 主持人镜头 | 10 000 | 9 886 | 98.86 |
| 主题字幕 | 10 000 | 9 676 | 96.76 |
| 静音 | 56 | 40 | 71.43 |

检测结果的样例如图 7~图 10 所示。在图 9 中所选样例的实际新闻段数为 16,但检测结果为 21,大于实际新闻段数,因为视频中有大于 0.3 s 的静音段并非发生新闻故事切换,所以导致表 2 中静音切分的准确率较低。图 10 中所选两段新闻故事样例的语义相似度较高可以选择进行合并。

图 7 主持人镜头检测结果样例

Fig. 7 Example of host lens detection results

图 8 主题字幕检测结果样例

Fig. 8 Example of subject caption detection results

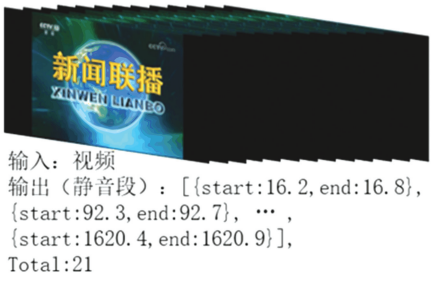


图 9 静音检测结果样例

Fig. 9 Example of mute detection results


图 10 语义相似分析结果样例

Fig. 10 Example of semantic similarity analysis results

该实验视频共有 683 个新闻故事,新闻故事分割检测实验结果见表 2。本算法共正确检出 672 个,新闻故事分割的平均查全率为 97.17%,查准率为 98.19%,平均 $F-measure$ 为 97.68%;2018~2020 年的新闻故事分割平均查全率均高于96.39%,其中 2020 年的平均查全率最高;3 年的新闻故事分割平均查准率均高于 97.84%,其中 2020 年的平均查准率最高。

表 2　新闻故事分割检测实验结果

Table 2　Experimental results of news story segmentation detection

| 视频编号 | 2018 年 | | 2019 年 | | 2015 年 | |
|---|---|---|---|---|---|---|
| | $R_r$ | $R_p$ | $R_r$ | $R_p$ | $R_r$ | $R_p$ |
| 1 | 98.31 | 97.52 | 82.56 | 100 | 97.69 | 100 |
| 2 | 96.95 | 100 | 98.14 | 100 | 100 | 100 |
| 3 | 92.81 | 100 | 97.65 | 100 | 98.11 | 97.38 |
| 4 | 100 | 100 | 98.19 | 98.38 | 100 | 97.33 |
| 5 | 95.10 | 95.63 | 100 | 95.32 | 100 | 98.23 |
| 6 | 100 | 96.83 | 100 | 97.45 | 100 | 100 |
| 7 | 100 | 100 | 100 | 98.23 | 92.44 | 96.99 |
| 8 | 100 | 100 | 100 | 100 | 94.86 | 97.10 |
| 9 | 100 | 98.07 | 92.76 | 100 | 100 | 100 |
| 10 | 100 | 97.52 | 95.05 | 92.36 | 98.06 | 100 |
| 11 | 82.14 | 97.73 | 100 | 98.93 | 100 | 100 |
| 12 | 96.87 | 95.52 | 100 | 98.60 | 92.52 | 94.53 |
| 13 | 92.38 | 96.51 | 97.61 | 98.03 | 99.82 | 97.31 |
| 14 | 98.48 | 100 | 98.31 | 97.76 | 98.99 | 96.62 |
| 15 | 92.86 | 100 | 94.23 | 92.54 | 100 | 100 |
| 总计 | 96.39 | 98.36 | 96.97 | 97.84 | 98.16 | 98.37 |

### 3.5　实验对比分析

本文分别与基于边界归类视频分割[5]和基于多风格探索视频分割(MSE-NSS)[14]进行对比实验。文献[14]通过字幕定位、字幕聚类和字幕跟踪的方法来判定故事边界；文献[5]基于字幕文本分类和主题字幕相似性来筛选故事边界，对比实验结果如图 11 所示。

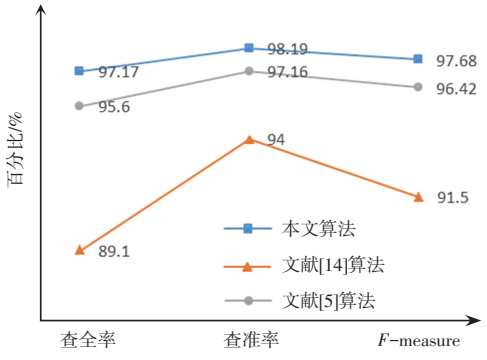


图 11　对比实验结果

Fig. 11　Comparison of experimental results

相较于基于边界归类视频分割和 MSE-NSS，本文的 $F$ - measure 高于两者。在主题字幕检测结果中，该算法的字幕检测准确率极高，能更好的对字幕进行定位、识别和追踪工作，基于边界归类视频分割通过提取字幕区域 SURF 特征点来计算 2 个镜头内的主题字幕帧匹配度，当字幕区域出现和主题字幕颜色相近的背景时可能会出现误判，易出现误检。在主持人镜头检测结果中，其检测率也极高，基于边界归类视频分割是基于主持人模板匹配来实现镜头标定，比较直方图差异来识别主持人镜头，当新闻中出现与主持人相似的镜头但又非主持人镜头时，容易误判。

因为本文算法结合了新闻视频的多种特性，所以在查全率和查准率上都有一定的提高，而分析两段新闻故事基本单元的语义相似性，决定两者是否属于同一新闻故事，进一步提升了视频分割的准确性。

## 4　结束语

新闻视频故事分割为新闻视频导航、检索等应用提供辅助，为新闻视频后续的处理打下基础。本文提出了一种基于多模态相似融合的新闻视频分割算法，该算法通过选定视频切割点获得候选新闻故事单元边界，利用新闻视频的视频特征、音频特征和语义特征等多模态特征获得最终的新闻故事。实验结果表明，本算法分割出的新闻故事具有很好的独立性，获得平均 97.17% 的查全率和 98.19% 的查准率，进一步提高了分割效果。本算法实现的新闻视频故事分割是针对特定的新闻视频，而面对较为复

（下转第 84 页）