(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110012349 B
(45)授权公告日 2019.09.20

(21)申请号 201910483232.X

(22)申请日 2019.06.04

(65)同一申请的已公布的文献号
申请公布号 CN 110012349 A

(43)申请公布日 2019.07.12

(73)专利权人 成都索贝数码科技股份有限公司
地址 610041 四川省成都市高新区新园南二路2号

(72)发明人 王炜 温序铭 谢超平 李杰 严照宇 孙翔 罗明利

(74)专利代理机构 成都弘毅天承知识产权代理有限公司 51230
代理人 蒋秀清

(51)Int.Cl.
*H04N 21/439*(2011.01)
*H04N 21/44*(2011.01)
*H04N 21/472*(2011.01)
*G06K 9/20*(2006.01)
*G06K 9/34*(2006.01)
*G10L 15/26*(2006.01)

(56)对比文件
CN 105868292 A,2016.08.17,全文.
CN 105844292 A,2016.08.10,全文.
CN 103902723 A,2014.07.02,全文.
US 2007296863 A1,2007.12.27,全文.
Pradip Panchal et al.Scene detection and retrieval of video using motion vector and occurrence rate of shot boundaries.《2012 NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING》.2012,
符茂胜等.视频结构化描述模型.《计算机应用》.2012,

审查员 赵斯曼

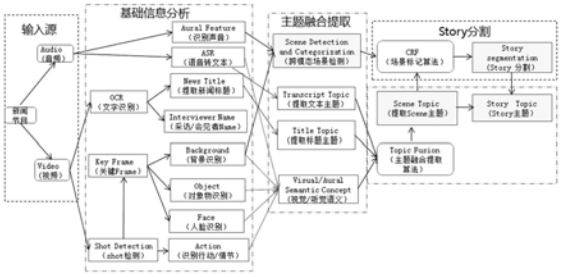
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种端到端的新闻节目结构化方法

(57)摘要

本发明公开了一种端到端的新闻节目结构化方法及其结构化框架体系,涉及新闻节目处理技术领域,本发明的方法包括对输入的新闻节目进行预处理,获取新闻节目的音频资源和视频资源;利用ASR语音识别技术、OCR文字识别技术和Shot Detection技术提取音频资源和视频资源内的基础信息;基于提取的基础信息,提取各模态的语义主题信息,并采用跨模态的主题融合提取算法,对各模态的语义主题信息进行融合聚类,输出Scene主题;同时进行跨模态场景检测,输出Scene层级;利用CRF场景标记算法对得到的Scene层级和Scene主题进行场景聚合和分割,输出具有相同语义的Story层级和Story主题,本发明重点关注具有明确语义含义的Story层和Scene层,便于新闻节目二次利用,提高了新闻节目的使用时效性。




CN 110012349 B

1.一种端到端的新闻节目结构化方法,其特征在于,包括如下步骤:

S1:对输入的新闻节目进行预处理,分别获取新闻节目的音频资源和视频资源;

S2:利用ASR语音识别技术、OCR文字识别技术和镜头检测技术提取音频资源和视频资源内的基础信息;

S3:基于S2提取的基础信息,提取各模态的语义主题信息,同时进行跨模态场景检测,并采用跨模态的主题融合提取算法,对各模态的语义主题信息进行融合聚类,输出Scene主题和输出Scene层级;

S4:利用CRF场景标记算法对S3中得到的Scene层级和Scene主题进行场景聚合和分割,输出具有相同语义的Story层级和Story主题;

其中,所述S2中对音频资源进行基础信息分析包括:

基于MFCCs音频特征的语音判定分析技术,识别音频资源的声音信息,通过音频特征分析判定语音播报的停顿间隔;

通过ASR语音识别技术将音频资源的语音内容转化为文本内容;

所述S2中对视频资源进行基础信息分析包括:

利用OCR文字识别技术对视频资源的文字部分进行文字识别,分析出文本信息,提取新闻节目标题;

利用镜头检测技术对视频资源的画面部分进行镜头检测,将新闻节目自底向上切分为若干具有相似视觉特征的镜头,并通过关键帧提取技术提取出所述具有相似视觉特征的镜头的关键帧序列,再根据所提取的关键帧序列对视频资源的背景、特定物体、人脸和行为进行识别;

所述S3具体包括如下步骤:

S3.1:基于ASR语音识别技术转化的文本内容和OCR文字识别技术提取的新闻节目标题,结合根据提取的关键帧序列得到的背景、特定物体、人脸和行为的识别信息,利用LDA无监督学习算法得到各模态的语义主题信息;

S3.2:以背景识别的时间点和基于MFCCs音频特征的语音判定分析技术得到的停顿间隔时间点作为跨模态场景检测的基线时间点,进行场景分割,对各模态的语义主题信息进行切分,输出Scene层级;

S3.3:采用跨模态的主题融合提取算法,对各场景的主题描述进行近似性计算,对主题相近的场景进行融合聚类,输出Scene主题。

# 一种端到端的新闻节目结构化方法

## 技术领域

[0001] 本发明涉及新闻节目处理技术领域，更具体的是涉及一种端到端的新闻节目结构化方法及其结构化框架体系。

## 背景技术

[0002] 随着时代的发展，技术的进步，视频的索引和检索是个重要的问题，并且具有重大意义。而电视新闻是视频中的一大部分，也是会被反复多次利用的视频。如电视新闻播出后的点播，需要将电视新闻流分段，然后再对每段电视新闻流进行元数据标注，从而快速进行索引和访问；电视新闻节目作为一种素材再次被利用，用作其他新闻节目的编辑材料，往往再次利用的是新闻的有价值片段，也需要将电视新闻流按照电视新闻结构进行分解，并对有利用价值的片段进行标注。

[0003] 新闻视频是视频的一种重要分支，他们包含着大量的有用信息，基于内容的视频检索系统指通过文本、图片或视频的其他特征在视频集中搜索需要的信息。

[0004] 一档新闻节目一般包括片头、主要内容介绍、新闻报道、天气预报及片尾，对于点播而言，需求则是对新闻报道(Story)这一层级进行索引和访问，对于作为素材再次利用即二次编辑而言，需求则是对Scene这一层级进行索引和访问；面对当前不断增加的海量新闻视频内容，使用原人工的方法进行新闻流分段和标注已经不可行，新闻节目的访问和二次编辑需要的实时性也得不到满足。

## 发明内容

[0005] 本发明的目的在于：为了解决使用原人工的方法进行不断增加的海量新闻流的分段和标注，新闻节目的访问和二次编辑的实时性得不到满足的问题，本发明提供一种端到端的新闻节目结构化方法及其结构化框架体系，综合了新闻语法、视觉特征、音频特征、文本语义等跨模态信息，融合采用计算机视觉、机器学习、自然语言处理等多种人工智能技术，一次性实现了新闻节目的Scene层级和Story层级结构切分和核心元数据自动描述。

[0006] 本发明为了实现上述目的具体采用以下技术方案：

[0007] 一种端到端的新闻节目结构化方法，包括如下步骤：

[0008] S1：对输入的新闻节目进行预处理，分别获取新闻节目的音频资源和视频资源；

[0009] S2：利用ASR语音识别技术、OCR文字识别技术和镜头检测技术提取音频资源和视频资源内的基础信息；

[0010] S3：基于S2提取的基础信息，提取各模态的语义主题信息，同时进行跨模态场景检测，并采用跨模态的主题融合提取算法，对各模态的语义主题信息进行融合聚类，输出Scene主题和输出Scene层级；

[0011] S4：利用CRF场景标记算法对S3中得到的Scene层级和Scene主题进行场景聚合和分割，输出具有相同语义的Story层级和Story主题。

[0012] 进一步的，所述S2中对音频资源进行基础信息分析包括：

[0013] 基于MFCCs音频特征的语音判定分析技术，识别音频资源的声音信息，通过音频特征分析判定语音播报的停顿间隔，音频的停顿间隔时间点将用于后续场景检测；

[0014] 通过ASR语音识别技术将音频资源的语音内容转化为文本内容，新闻节目中播音员的语音播报内容对于理解新闻节目语义含义、元数据提取都非常重要，因此语音识别技术的分析工作是基础分析工作。

[0015] 进一步的，所述S2中对视频资源进行基础信息分析包括：

[0016] 利用OCR文字识别技术对视频资源的文字部分进行文字识别，分析出文本信息，提取新闻节目标题；

[0017] 利用镜头检测技术对视频资源的画面部分进行Shot检测，将新闻节目自底向上切分为若干具有相似视觉特征的镜头，并通过关键帧提取技术提取出所述具有相似视觉特征的镜头的关键帧序列，再根据所提取的关键帧序列对视频资源的背景、特定物体、人脸和行为进行识别，这些识别信息将用于后续场景检测、主题融合分析和元数据自动填写流程环节。

[0018] 进一步的，所述S3具体包括如下步骤：

[0019] S3.1：基于ASR语音识别技术转化的文本内容和OCR文字识别技术提取的新闻节目标题，结合根据提取的关键帧序列得到的背景、特定物体、人脸和行为的识别信息，利用LDA无监督学习算法得到各模态的语义主题信息，这些语义主题信息可看作各模态对当前视频片段的内容理解的概要表达，但这些独立模态的表达可能是不准确的，有缺失的，甚至是错误的，所以，我们还需要通过一种新闻节目多模态融合算法将各模态的主题表达进行融合聚类，最终形成相对正确的主题概要表达；

[0020] S3.2：由于新闻节目视频画面是基础，同一个场景中不论镜头如何切换，其画面的背景是相同或接近的，因此，以背景识别的时间点和基于MFCCs音频特征的语音判定分析技术得到的停顿间隔时间点作为跨模态场景检测的基线时间点，进行场景分割，对各模态的语义主题信息进行切分，输出Scene层级；

[0021] S3.3：采用跨模态的主题融合提取算法，对各场景的主题描述进行近似性计算，对主题相近的场景进行融合聚类，输出Scene主题。

[0022] 进一步的，通过前面步骤基本完成了新闻节目各场景的切分和主题，人物，关键词等核心元数据的自动提取，但是还需将这些场景准确的组合成具备完整故事的节目片段。所以，我们采用基于CRF算法通过对一定样本数据进行学习，将若干场景分割和聚合为不同的Story片段中。CRF算法输入是一组Scene序列的视觉类别特征和文本主题特征，输出是对各场景序列的位置标签。这些位置标签将可用于切分和组合Story片段，利用CRF场景标记算法对S3中得到的Scene层级和Scene主题进行场景聚合和分割，输出相同语义的Story，构成Story层级和Story主题。

[0023] 本发明的有益效果如下：

[0024] 1、本发明对不同来源的新闻节目，通过多维度结合ASR语音识别技术、OCR文字识别技术等，进行跨模态的特征融合，主题融合提取，再基于CRF场景标记算法，提高了Story分割及Scene和Story主题提取的准确率，同时获取到的Story层级、Scene层级，方便新闻节目的点播和二次编辑直接取用，提高了使用时效性，端到端的整个过程系统自动完成，有效避免了人为干扰信息，减少出错，同时节省时间。

[0025] 2、本发明对不同来源的新闻节目，充分利用其视频、文字、语音的特征信息，通过各智能识别分析技术分析出基础信息，采用跨模态的特征融合，对主题进行融合，形成Scene主题和层级，再基于CRF场景标记算法，实现Story分割，产生结构化体系中的具有完整故事描述的节目片段 Story层次及Story主题，既充分利用了各种来源视频、文字、语音的特征信息，又有效避免了干扰信息，确保提取结果的精准性。

## 附图说明

[0026] 图1是本发明的新闻节目结构化方法流程示意图。

[0027] 图2是本发明的新闻节目结构化框架体系示意图。

## 具体实施方式

[0028] 为了本技术领域的人员更好的理解本发明，下面结合附图和以下实施例对本发明作进一步详细描述。

[0029] 实施例1

[0030] 如图1和图2所示，本实施例提供一种端到端的新闻节目结构化方法，包括如下步骤：

[0031] S1：对输入的新闻节目进行预处理，分别获取新闻节目的音频资源和视频资源；

[0032] S2：利用ASR语音识别技术、OCR文字识别技术和镜头检测技术提取音频资源和视频资源内的基础信息；

[0033] 所述S2中对音频资源进行基础信息分析包括：

[0034] 基于MFCCs音频特征的语音判定分析技术，识别音频资源的声音信息，通过音频特征分析判定语音播报的停顿间隔，音频的停顿间隔时间点将用于后续场景检测；

[0035] 通过ASR语音识别技术将音频资源的语音内容转化为文本内容，新闻节目中播音员的语音播报内容对于理解新闻节目语义含义、元数据提取都非常重要，因此语音识别技术的分析工作是基础分析工作；

[0036] 所述S2中对视频资源进行基础信息分析包括：

[0037] 利用OCR文字识别技术对视频资源的文字部分进行文字识别，分析出文本信息，提取新闻节目标题和与会者名字信息；

[0038] 利用镜头检测技术对视频资源的画面部分进行Shot检测，将新闻节目自底向上切分为若干具有相似视觉特征的镜头，并通过关键帧提取技术提取出所述具有相似视觉特征的镜头的关键帧序列，再基于CNN、GAN、C3D等深度神经网络模型根据所提取的关键帧序列对视频资源的背景、特定物体、人脸和行为进行识别，这些识别信息将用于后续场景检测、主题融合分析和元数据自动填写流程环节；

[0039] S3：基于S2提取的基础信息，提取各模态的语义主题信息，同时进行跨模态场景检测，并采用跨模态的主题融合提取算法，对各模态的语义主题信息进行融合聚类，输出Scene主题和输出Scene层级，具体包括如下步骤：

[0040] S3.1：基于ASR语音识别技术转化的文本内容和OCR文字识别技术提取的新闻节目标题，结合根据提取的关键帧序列得到的背景、特定物体、人脸和行为的识别信息，利用LDA无监督学习算法得到各模态的语义主题信息，这些语义主题信息可看作各模态对当前视频

片段的内容理解的概要表达，但这些独立模态的表达可能是不准确的，有缺失的，甚至是错误的，所以，我们还需要通过一种新闻节目多模态融合算法将各模态的主题表达进行融合聚类，最终形成相对正确的主题概要表达；

[0041] S3.2：本实施例中新闻节目结构化最小单元是Scene(场景)，因此场景的精准检测定位尤为重要，由于新闻节目视频画面是基础，同一个场景中不论镜头如何切换，其画面的背景是相同或接近的，因此，以背景识别的时间点和基于MFCCs音频特征的语音判定分析技术得到的停顿间隔时间点作为跨模态场景检测的基线时间点，进行场景分割，对各模态的语义主题信息进行切分，输出Scene层级，可忽略掉一些视觉场景错误切分的时间点；

[0042] S3.3：采用跨模态的主题融合提取算法，对各场景的主题描述进行近似性计算，对主题相近的场景进行融合聚类，输出Scene主题；

[0043] S4：利用CRF场景标记算法对S3中得到的Scene层级和Scene主题进行场景聚合和分割，输出具有相同语义的Story层级和Story主题，具体为：

[0044] 通过前面步骤基本完成了新闻节目各场景的切分和主题、人物、关键词等核心元数据的自动提取，但是还需将这些场景准确的组合成具备完整故事的节目片段；所以，我们采用基于CRF算法通过对一定样本数据进行学习，将若干场景分割和聚合为不同的Story片段中。CRF算法的输入是一组Scene序列的视觉类别特征和文本主题特征，输出的是对各场景序列的位置标签。这些位置标签将可用于切分和组合Story片段，即利用CRF场景标记算法对S3中得到的Scene层级和Scene主题进行场景聚合和分割，输出相同语义的Story，构成Story层级和Story主题。

[0045] 如图2所示，本实施例在跨模态场景检测后输出Scene层级，主题融合提取后输出Scene主题，然后经过CRF场景标记算法进行Story分割后输出Story层级以及Story主题，由于Shot层级和Frame帧在新闻节目中的独立语义信息不够丰富，因此在本实施例中并不对其进行过多处理，本实施例重点关注具有明确语义含义的Story层和Scene层，通过OCR、ASR等技术初始化信息解析，找出Scene的主题、分类、人物、关键字等信息，经过提取主题、融合等复杂处理，输出Scene层、Scene主题、Story层及Story主题，多个Shot组成Scene，Scene作为素材被二次编辑使用；多个Scene构成Story，电视新闻的点播可直接使用Story层级，经过端到端的新闻节目结构化处理，避免了人工操作的繁琐和出错，提高了新闻节目使用时效性。

[0046] 以上所述，仅为本发明的较佳实施例，并不用以限制本发明，本发明的专利保护范围以权利要求书为准，凡是运用本发明的说明书及附图内容所作的等同结构变化，同理均应包含在本发明的保护范围内。

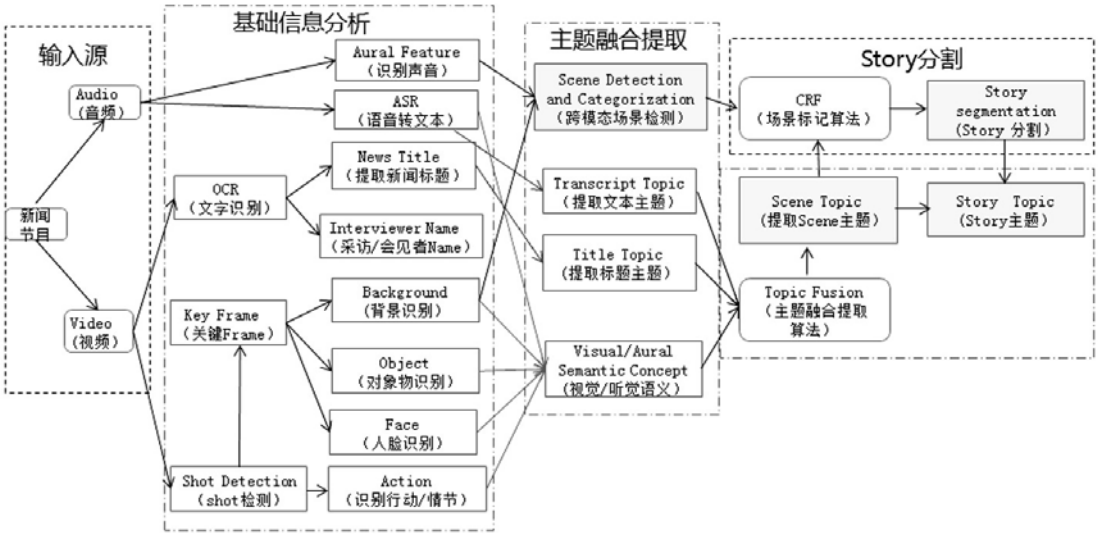

输入源
Audio
（音频）
新闻
节目
Video
（视频）
基础信息分析
Aural Feature
（识别声音）
ASR
（语音转文本）
OCR
（文字识别）
News Title
（提取新闻标题）
Interviewer Name
（采访/会见者Name）
Key Frame
（关键Frame）
Background
（背景识别）
Object
（对象物识别）
Face
（人脸识别）
Shot Detection
（shot检测）
Action
（识别行动/情节）
主题融合提取
Scene Detection
and Categorization
（跨模态场景检测）
Transcript Topic
（提取文本主题）
Title Topic
（提取标题主题）
Visual/Aural
Semantic Concept
（视觉/听觉语义）
Story分割
CRF
（场景标记算法）
Story
segmentation
（Story 分割）
Scene Topic
（提取Scene主题）
Story Topic
（Story主题）
Topic Fusion
（主题融合提取
算法）

图1

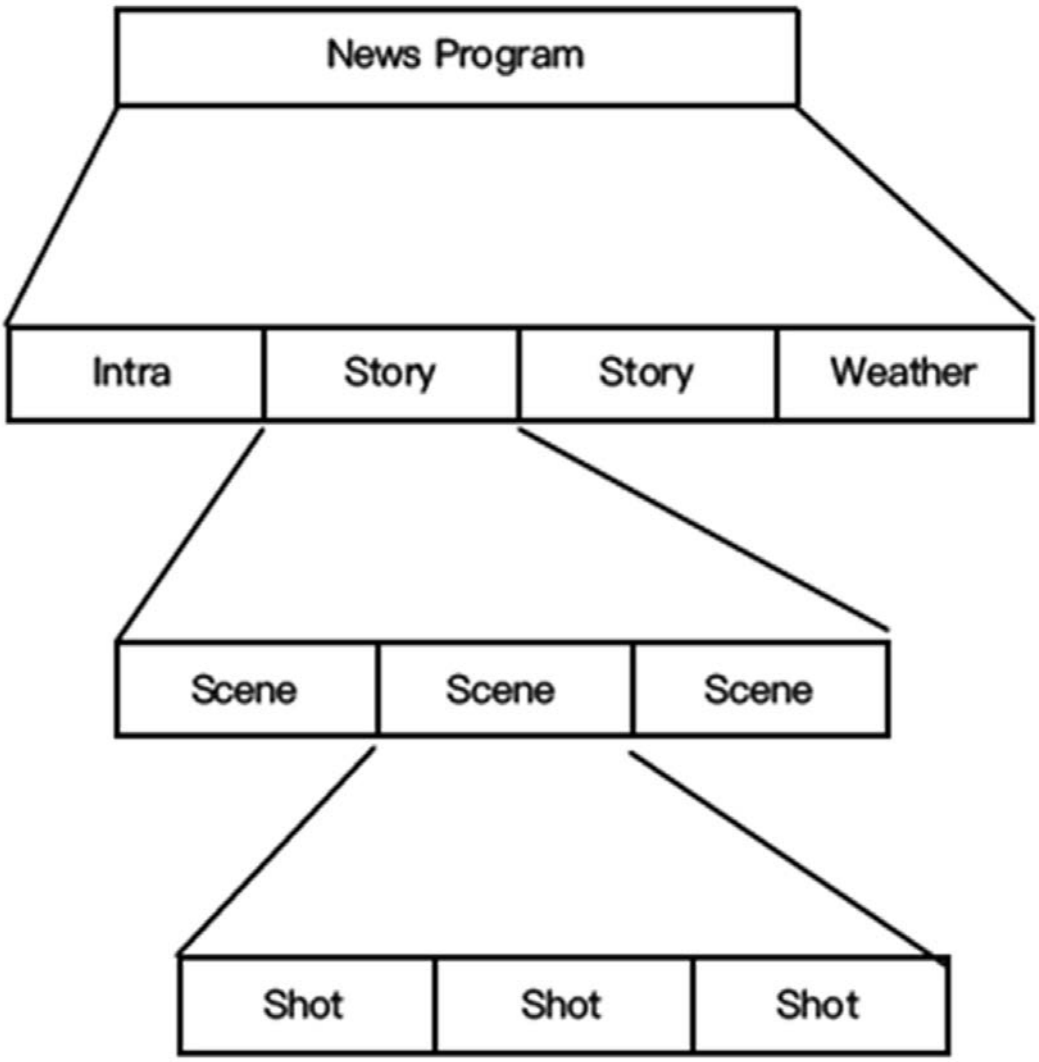

News Program
Intra
Story
Story
Weather
Scene
Scene
Scene
Shot
Shot
Shot

图2