(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114005079 A
(43)申请公布日 2022.02.01

(21)申请号 202111666523.6

(22)申请日 2021.12.31

(71)申请人 北京金茂教育科技有限公司
地址 100000 北京市丰台区汽车博物馆东路8号院3号楼15层1503

(72)发明人 赵悦汐 程红兵 鞠剑伟 昝晨辉

(74)专利代理机构 北京中索知识产权代理有限公司 11640
代理人 葛靖

(51)Int.Cl.
G06V 20/40(2022.01)
G06V 40/20(2022.01)
G06V 40/16(2022.01)
G06V 30/10(2022.01)
H04L 65/60(2022.01)

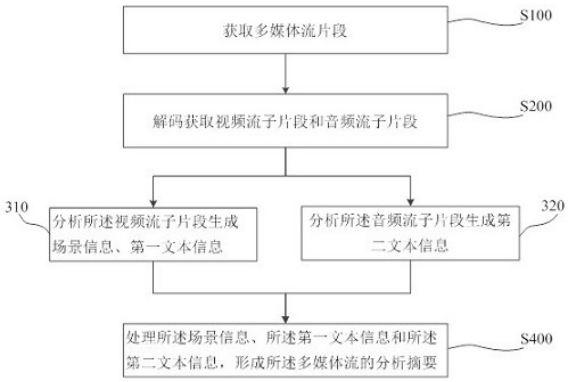
权利要求书2页 说明书10页 附图1页

(54)发明名称

多媒体流处理方法及装置

(57)摘要

本申请提供一种多媒体流处理方法及装置。其中,所述方法包括:获取多媒体流片段;解码获取视频流子片段和音频流子片段;分析所述视频流子片段生成场景信息、第一文本信息;分析所述音频流子片段生成第二文本信息;处理所述场景信息、所述第一文本信息和所述第二文本信息,形成所述多媒体流的分析摘要。通过将多媒体流文件进行拆解,能够通过有效结合各种独立的AI模块进行复杂场景下多媒体文件内容识别,有效提升了复杂场景下现有独立的AI技术识别效率。



CN 114005079 A

1.一种多媒体流处理方法,其特征在于,包括以下步骤:

获取多媒体流片段;

解码获取视频流子片段和音频流子片段;

分析所述视频流子片段生成场景信息、第一文本信息;

分析所述音频流子片段生成第二文本信息;

处理所述场景信息、所述第一文本信息和所述第二文本信息,形成所述多媒体流的分析摘要。

2.如权利要求1所述的多媒体流处理方法,其特征在于,分析所述视频流子片段生成场景信息,具体包括:

分析视频流子片段,生成面向对象的身份特征识别信息和对对象动作行为的描述信息。

3.如权利要求1所述的多媒体流处理方法,其特征在于,分析所述视频流子片段生成第一文本信息,具体包括:

分析视频流子片段,生成对象动作行为指向的第一文本信息。

4.如权利要求3所述的多媒体流处理方法,其特征在于,分析视频流子片段,生成对象动作行为指向的第一文本信息,具体包括:

分析视频流子片段,获取持续预设时长的图像;

使用OCR对所述图像进行识别,生成第一文本信息。

5.如权利要求4所述的多媒体流处理方法,其特征在于,所述第一文本信息至少包括教学环节信息和知识点信息其中之一。

6.如权利要求1所述的多媒体流处理方法,其特征在于,所述第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。

7.如权利要求1所述的多媒体流处理方法,其特征在于,处理所述场景信息、所述第一文本信息和所述第二文本信息,形成所述多媒体流的分析摘要,具体包括:

对场景信息、第一文本信息和第二文本信息进行交叉验证,形成所述多媒体流的分析摘要。

8.一种多媒体流处理装置,其特征在于,包括:

获取模块,用于获取多媒体流片段;

解码模块,用于解码获取视频流子片段和音频流子片段;

视频分析模块,用于分析所述视频流子片段生成场景信息、第一文本信息;

音频分析模块,用于分析所述音频流子片段生成第二文本信息;

分析摘要生成模块,用于处理所述场景信息、所述第一文本信息和所述第二文本信息,形成所述多媒体流的分析摘要。

9.如权利要求8所述的多媒体流处理装置,其特征在于,所述视频分析模块用于分析所述视频流子片段生成场景信息,具体用于:

分析视频流子片段,生成面向对象的身份特征识别信息和对对象动作行为的描述信息。

10.如权利要求8所述的多媒体流处理装置,其特征在于,所述视频分析模块用于分析所述视频流子片段生成第一文本信息,具体用于:

分析视频流子片段,生成对象动作行为指向的第一文本信息。

多媒体流处理方法及装置

技术领域

[0001] 本申请涉及多媒体信息识别技术领域,尤其涉及一种多媒体流处理方法及装置。

背景技术

[0002] 随着AI技术的持续发展和普及,市场上出现了很多成熟的AI模块,比如阿里多媒体AI,可以用来处理媒体中的信息流。例如,多媒体中的视频流、音频流,或视频流与音频流结合的信息流。在这些多媒体流处理的过程中,可以通过相应的AI模块对获取的多媒体文件中相应内容进行识别。

[0003] 在实现现有技术的过程中,发明人发现:

目前常见的AI模块识别方式单一。面对复杂的待识别场景,无法通过单一的AI模块进行分析,从而降低了多媒体文件的识别效率。

[0004] 因此,需要提供一种多媒体流处理方法及装置,用以解决复杂场景下现有独立的AI技术识别效率低的技术问题。

发明内容

[0005] 本申请实施例提供一种多媒体流处理方法及装置,用以解决复杂场景下现有独立的AI技术识别效率低的技术问题。

[0006] 具体的,一种多媒体流处理方法,包括以下步骤:

获取多媒体流片段;

解码获取视频流子片段和音频流子片段;

分析所述视频流子片段生成场景信息、第一文本信息;

分析所述音频流子片段生成第二文本信息;

处理所述场景信息、所述第一文本信息和所述第二文本信息,形成所述多媒体流的分析摘要。

[0007] 进一步的,分析所述视频流子片段生成场景信息,具体包括:

分析视频流子片段,生成面向对象的身份特征识别信息和对对象动作行为的描述信息。

[0008] 进一步的,分析所述视频流子片段生成第一文本信息,具体包括:

分析视频流子片段,生成对象动作行为指向的第一文本信息。

[0009] 进一步的,分析视频流子片段,生成对象动作行为指向的第一文本信息,具体包括:

分析视频流子片段,获取持续预设时长的图像;

使用OCR对所述图像进行识别,生成第一文本信息。

[0010] 进一步的,所述第一文本信息至少包括教学环节信息和知识点信息其中之一。

[0011] 进一步的,所述第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。

[0012] 进一步的，处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要，具体包括：

对场景信息、第一文本信息和第二文本信息进行交叉验证，形成所述多媒体流的分析摘要。

[0013] 本申请实施例还提供一种多媒体流处理装置。

[0014] 具体的，一种多媒体流处理装置，包括：

获取模块，用于获取多媒体流片段；

解码模块，用于解码获取视频流子片段和音频流子片段；

视频分析模块，用于分析所述视频流子片段生成场景信息、第一文本信息；

音频分析模块，用于分析所述音频流子片段生成第二文本信息；

分析摘要生成模块，用于处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要。

[0015] 进一步的，所述视频分析模块用于分析所述视频流子片段生成场景信息，具体用于：

分析视频流子片段，生成面向对象的身份特征识别信息和对对象动作行为的描述信息。

[0016] 进一步的，所述视频分析模块用于分析所述视频流子片段生成第一文本信息，具体用于：

分析视频流子片段，生成对象动作行为指向的第一文本信息。

[0017] 通过申请实施例提供的技术方案，至少具有如下有益效果：

通过将多媒体流文件进行拆解，能够通过有效结合各种独立的AI模块进行复杂场景下多媒体文件内容识别，有效提升了复杂场景下现有独立的AI技术识别效率。

附图说明

[0018] 此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图1为本申请实施例提供的一种多媒体流处理方法的流程图。

[0019] 图2为本申请实施例提供的一种多媒体流处理装置的结构示意图。

[0020] 100 多媒体流处理装置

11 获取模块

12 解码模块

13 视频分析模块

14 音频分析模块

15 分析摘要生成模块。

具体实施方式

[0021] 为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做

出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0022] 可以理解的是，多媒体流文件记录有视频流信息以及音频流信息。其中，所述视频流信息主要对应多媒体文件中的连续若干帧图像；所述音频流信息对应为多媒体文件中的语音信息合集。由此可知，视频流信息对应记录有与环境相关的场景信息以及文本信息；音频流信息对应记录有与视频流中相关环境对应的语音信息。这里的场景信息可以理解为每一帧图像中记录的、与呈现对象相关的对象信息；这里的文本信息可以理解为每一帧图像中记录的与文字相关的符号信息。

[0023] 通过单一的AI模块，可以对所述视频流中的场景信息或文本信息，或音频流信息进行单一的识别，从而识别出视频文件中的对象行为或存在的文本信息，或音频文件中的语音内容。但是复杂场景下的多媒体文件，往往既包含视频信息，又包含音频信息。若继续通过单一的AI模块进行识别，则无法全面识别到多媒体文件中记录的视频流信息以及音频流信息。这样，使得识别到的目标内容与多媒体文件记录的真实内容存在一定的误差。虽然也可以利用若干不同的单一AI模块进行同时进行复杂场景下记录内容的识别，但是每一单一AI模块的计算工作量较大。这样，降低了多媒体文件的识别速度，且不利于进行相关识别结果的结构化合并。

[0024] 本申请实施例提供一种多媒体流处理方法，主要用于处理复杂场景下的多媒体文件。在本申请提供的一种具体实施方式中，所述多媒体流处理方法可以用于处理记录有课堂教学过程这一复杂场景的多媒体文件。具体的，请参照图1，一种多媒体流处理方法，包括以下步骤：

S100：获取多媒体流片段。

[0025] 这里的多媒体流片段可以理解为记录有相应场景的文字、图形、影像、动画、声音及视频等媒体信息的文件。在本申请提供的一种具体实施方式中，获取的多媒体流片段为具有一定时长，且记录有课堂教学场景的多媒体文件。所述多媒体流片段可以通过相应的视频拍摄设备拍摄得到。这样，能够对课堂实时场景进行拍摄，从而得到记录有课堂教学过程中的声音、文字、图片、人员对象等信息的多媒体文件。

[0026] S200：解码获取视频流子片段和音频流子片段。

[0027] 这里的视频流子片段可以理解为多媒体片段中的图像信息。这里的音频流子片段可以理解为多媒体片段中的声音信息。对获取的过媒体流片段进行解码，即将具有一定时长的多媒体文件中的图像信息以及声音信息提取出来，并转换为预设文件格式的连续若干帧图像以及连续音频，从而得到多媒体流片段对应的视频流子片段和音频流子片段。

[0028] 当获取的多媒体流片段为具有一定时长，且记录有课堂教学场景的多媒体文件时，经解码，对应得到记录有课堂教学过程中文字、图片、人员对象等信息的视频流子片段，以及记录有课堂教学过程中声音信息的音频流子片段。

[0029] S310：分析所述视频流子片段生成场景信息、第一文本信息。

[0030] 可以理解的是，多媒体流片段中的连续若干帧图像即可构成视频流子片段。而每一帧图像均记录有相应的场景信息。在课堂教学场景中，所述视频流子片段为记录有课堂教学过程中文字、图片、人员对象等信息连续若干帧图像。经具有相应功能的AI模块分析，即可得到当前视频流子片段中人员对象对应的具体场景信息，以及当前视频流子片段中文字、图片对应的具体文本信息。

[0031] 具体的，通过对当前视频流子片段中人员对象的识别，能够确定当前人员对象的具体动作类别，从而便于确定当前视频流子片段对应的具体课堂教学场景。通过对当前视频流子片段中文字、图片对应的具体文本信息的识别，能够确定当前视频流子片段对应的文字、图片信息对应的具体文本种类或描述内容，从而能够生成当前视频流子片段对应的第一文本信息。这里的第一文本信息可以理解为根据视频流子片段生成的文本信息。

[0032] S320：分析所述音频流子片段生成第二文本信息。

[0033] 可以理解的是，根据多媒体流片段中的语音信息即可生成音频流子片段。在课堂教学场景中，所述音频流子片段为记录有课堂教学过程中声音信息相关的文件。具有相应功能的AI模块分析，即可确定所述音频流子片段的具体讲述内容，并得到与所述音频流子片段讲述内容相对应的第二文本信息。这里的第二文本信息可以理解为根据音频流子片段生成的文本信息。

[0034] S400：处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要。

[0035] 这里的分析摘要可以理解为当前处理的多媒体流片段对应的具体实时场景的概述。对场景信息、第一文本信息和第二文本信息进行处理，主要是识别其中与多媒体流片段对应的实时场景关联度较高的重点数据。通过对识别到的目标数据进行整合，即可得到当前处理的多媒体流片段对应的具体教学过程。将多媒体流片段拆分为视频流子片段以及音频流子片段，并通过具有相应功能的AI模块分析，能够有效减小功能单一的AI模块的处理数据量，并能够准确选取具有相应分析功能的分析模块，从而提高了多媒体流片段的识别效率。

[0036] 进一步的，在本申请提供的一种优选实施方式中，分析所述视频流子片段生成场景信息，具体包括：分析视频流子片段，生成面向对象的身份特征识别信息和对对象动作行为的描述信息。

[0037] 这里对象的身份特征识别信息可以理解为对象的脸部特征信息。可以理解的是，获取带有某一人员对象脸部的图像，通过经预训练的识别算法对其脸部特征进行识别，即可确定该人员对象的具体身份信息。例如，确定某一学生的姓名、学号等信息，或某一教师的姓名、工号等信息。

[0038] 这里对对象动作行为的描述信息可以理解为当前视频流子片段中，人员对象的具体动作类别。可以理解的是，获取带有某一人员对象躯体动作的图像，经预训练的识别算法对该图像进行识别，即可确定该人员对象的具体动作类别。例如，确定该人员对象当前动作行为为书写行为、起立行为或板书行为等。

[0039] 通过识别视频流子片段中涉及对象的具体身份信息以及行为信息，可以确定视频流子片段对应教学场景中的师生行为，从而提升了多媒体流分析摘要的准确率。

[0040] 进一步的，在本申请提供的一种优选实施方式中，分析所述视频流子片段生成第一文本信息，具体包括：分析视频流子片段，生成对象动作行为指向的第一文本信息。

[0041] 这里的对象动作行为指向的第一文本信息，可以理解为与人员对象的具体行为具有一定关联度的文本信息。可以理解的是，根据视频流子片段生成的第一文本信息中包括课堂教学场景下的文字、图片信息对应的具体文本种类或描述内容。例如，课堂中的PPT展示页、教室背景相关的黑板报信息。在课堂教学场景中，所述课堂中的PPT展示页中的文字

信息，即为与人员对象动作行为相关的文本信息。但是，黑板报信息为视频流子片段中的背景信息，与当前所处场景下的人员动作行为无关，则不属于第一文本信息。

[0042] 通过针对性分析视频流子片段中与人员对象动作相关的第一文本信息，可以减少相应功能模块的数据处理量，同时增加了相应功能模块识别的识别准确度，从而有效提升了第一文本信息的分析效率。

[0043] 进一步的，在本申请提供的一种优选实施方式中，分析视频流子片段，生成对象动作行为指向的第一文本信息，具体包括：分析视频流子片段，获取持续预设时长的图像；使用OCR对所述图像进行识别，生成第一文本信息。

[0044] 可以理解的是，与对象动作行为指向第一文本信息为与当前视频流子片段对应场景下，与对象行为具有一定关联度的文本信息。在课堂教学场景下，与对象行为具有一定关联度的文本信息，可以理解为教学内容相关的文本信息。例如，教师书写的板书、PPT展示页中的文字信息等。需要指出的是，在实际教学过程中，若对象动作行为指向的文本信息较为重要，则人员对象就相应文本信息展开相关行为时，会持续一定的时长。即，具有重要文本信息的图像停留时间较长。而对象动作行为指向的文本信息重要度较低或为无价值的文本信息时，则相应具有较短的时长。即，具有非重要文本信息的图像停留时间较短。因此，可以根据图像持续时长判断图像中文本的重要程度。这里的图像持续时长可以根据实际情况或者经验值预设。若某帧图像持续时长满足预设条件，即可展开后续的文本识别过程。对应的，若某帧图像持续时长不满足预设条件，说明其对应的文本重要性较低，无需展开相应的识别。这样，能够在保证第一文本识别准确度的基础上，增加第一文本的识别效率。

[0045] 具体的，对满足持续预设时长的图像记性文本信息识别时，可以通过OCR识别方式。即，利用光学字符识别技术，识别视频流子片段中满足预设条件的图像中的字符信息。例如，通过OCR识别方式将课堂展示PPT中满足预设条件的PPT页中的相关内容识别为第一文本信息。或者，通过OCR识别方式将课堂黑板上特定标记的文字内容识别为第一文本信息。

[0046] 进一步的，在本申请提供的一种优选实施方式中，所述第一文本信息至少包括教学环节信息和知识点信息其中之一。

[0047] 这里的教学环节信息可以理解为视频流子片段中满足预设时长的图像中，记录的当前图像对应的环节。例如，PPT中当前也对应的具体标题。这里的知识点信息可以理解为视频流子片段中满足预设时长的图像中，记录的当前图像特定的标注内容。

[0048] 可以理解的是，课堂教学过程中的相关板书资料，不可避免会存在一些与教学内容无关的信息。这里将其称为无价值信息。这些无价值信息的识别，在增加计算量的同时，还会增加第一文本中无意义内容的占比。因此，通过针对性地识别视频流子片段中教学环节信息或知识点信息，便于得到更为精准的第一文本标识信息，降低了第一文本信息的冗余度，从而便于进行多媒体分析摘要的确定。

[0049] 进一步的，在本申请提供的一种优选实施方式中，所述第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。

[0050] 可以理解的是，第二文本信息对应为根据音频流子片段生成的文本信息。在课堂教学场景中，所述音频流子片段为记录有课堂教学过程中与声音信息相关的文件。在实际应用中，将音频流子片段识别为第二文本，可以用于制作与视频流子片段同步的字幕信息。

过自然语言处理模型，能够对字幕中的文字信息进行纠错，并且还能进行关键词信息、提问信息、感情描述信息的提取。这些信息均可以理解为多媒体流分析摘要的形成要素。因此，经分析得到的第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。这样，能够保证多媒体流分析摘要的准确性。

[0051] 进一步的，在本申请提供的一种优选实施方式中，处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要，具体包括：对场景信息、第一文本信息和第二文本信息进行交叉验证，形成所述多媒体流的分析摘要。

[0052] 这里所述的交叉验证可以理解为将分析得到的场景信息、第一文本信息和第二文本信息等各类数据归总到不同的课程结构中。可以理解的是，进行场景信息、第一文本信息的分析，是根据视频流子片段完成的。进行第二文本信息的识别，是根据音频流子片段完成的。所述视频流子片段与音频流子片段均为从多媒体流片段中提取得到的。若仅根据场景信息、第一文本信息和第二文本信息中的某一信息进行多媒体流的分析摘要，无法得到具有较高准确性的多媒体流分析摘要。因此，需要综合考虑场景信息、第一文本信息和第二文本信息，并将其根据课程结构进行相应的归类汇总，从而得到准确性较高的多媒体流分析摘要。这一过程，也可理解为根据获取的多媒体视频内容，完成视频内容数据的拆解，并将拆解数据重新归类关联到相应的课堂环节。

[0053] 在本申请提供的一种具体实施方式中，可以将课堂教学场景中的将课程内容拆解为教学内容、师生行为、师生语言三大类。其中，教学内容可以主要体现在教学PPT及老师讲课的语音内容中；师生行为主要体现在动作肢体的行为变化上；师生语言主要体现在语言交流上。因此，针对课程内容，利用OCR识别技术进行PPT/黑板上的特定标记的文字内容的识别。这时，完成了课程总体结构的第一次划分。即将课堂环节进行了区分。通过人脸识别技术并结合对动作行为的识别，能够将老师授课、学生板书、举手、起立、阅读、书写等一系列行为动作进行行为划分。最后，再将课堂中的语音进行实时字幕翻译，并通过自然语言处理能力，能够将字幕中的文字信息纠错，并且还能提取其中与关键词信息、提问信息、情感描述信息等相关的语境信息。这样，得到了相关的场景信息、第一文本信息与第二文本信息。将得到的场景信息、第一文本信息与第二文本信息等各类数据归总到不同的课程结构中，即可完成对课堂教学视频内容的拆解，并将拆解得到的相关数据重新归类关联到相应的课堂的环节。

[0054] 本申请实施例还提供一种多媒体流处理装置100，主要用于处理复杂场景下的多媒体文件。在本申请提供的一种具体实施方式中，所述多媒体流处理装置100可以用于处理记录有课堂教学过程这一复杂场景的多媒体文件。具体的，请参照图2，一种多媒体流处理装置，包括：

获取模块11，用于获取多媒体流片段；

解码模块12，用于解码获取视频流子片段和音频流子片段；

视频分析模块13，用于分析所述视频流子片段生成场景信息、第一文本信息；

音频分析模块14，用于分析所述音频流子片段生成第二文本信息；

分析摘要生成模块15，用于处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要。

[0055] 获取模块11，用于获取多媒体流片段。这里的多媒体流片段可以理解为记录有相

应场景的文字、图形、影像、动画、声音及视频等媒体信息的文件。在本申请提供的一种具体实施方式中，获取的多媒体流片段为具有一定时长，且记录有课堂教学场景的多媒体文件。所述多媒体流片段可以通过相应的视频拍摄设备拍摄得到。这样，能够对课堂实时场景进行拍摄，从而得到记录有课堂教学过程中的声音、文字、图片、人员对象等信息的多媒体文件。

[0056] 解码模块12，用于解码获取视频流子片段和音频流子片段。这里的视频流子片段可以理解为多媒体片段中的图像信息。这里的音频流子片段可以理解为多媒体片段中的声音信息。对获取的过媒体流片段进行解码，即将具有一定时长的多媒体文件中的图像信息以及声音信息提取出来，并转换为预设文件格式的连续若干帧图像以及连续音频，从而得到多媒体流片段对应的视频流子片段和音频流子片段。

[0057] 当获取的多媒体流片段为具有一定时长，且记录有课堂教学场景的多媒体文件时，经解码，对应得到记录有课堂教学过程中文字、图片、人员对象等信息的视频流子片段，以及记录有课堂教学过程中声音信息的音频流子片段。

[0058] 视频分析模块13，用于分析所述视频流子片段生成场景信息、第一文本信息。可以理解的是，多媒体流片段中的连续若干帧图像即可构成视频流子片段。而每一帧图像均记录有相应的场景信息。在课堂教学场景中，所述视频流子片段为记录有课堂教学过程中文字、图片、人员对象等信息连续若干帧图像。经具有相应功能的AI模块分析，即可得到当前视频流子片段中人员对象对应的具体场景信息，以及当前视频流子片段中文字、图片对应的具体文本信息。

[0059] 具体的，通过对当前视频流子片段中人员对象的识别，能够确定当前人员对象的具体动作类别，从而便于确定当前视频流子片段对应的具体课堂教学场景。通过对当前视频流子片段中文字、图片对应的具体文本信息的识别，能够确定当前视频流子片段对应的文字、图片信息对应的具体文本种类或描述内容，从而能够生成当前视频流子片段对应的第一文本信息。这里的第一文本信息可以理解为根据视频流子片段生成的文本信息。

[0060] 音频分析模块14，用于分析所述音频流子片段生成第二文本信息。可以理解的是，根据多媒体流片段中的语音信息即可生成音频流子片段。在课堂教学场景中，所述音频流子片段为记录有课堂教学过程中声音信息相关的文件。具有相应功能的AI模块分析，即可确定所述音频流子片段的具体讲述内容，并得到与所述音频流子片段讲述内容相对应的第二文本信息。这里的第二文本信息可以理解为根据音频流子片段生成的文本信息。

[0061] 分析摘要生成模块15，用于处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要。这里的分析摘要可以理解为当前处理的多媒体流片段对应的具体实时场景的概述。对场景信息、第一文本信息和第二文本信息进行处理，主要是识别其中与多媒体流片段对应的实时场景关联度较高的重点数据。通过对识别到的目标数据进行整合，即可得到当前处理的多媒体流片段对应的具体教学过程。将多媒体流片段拆分为视频流子片段以及音频流子片段，并通过具有相应功能的AI模块分析，能够有效减小功能单一的AI模块的处理数据量，并能够准确选取具有相应分析功能的分析模块，从而提高了多媒体流片段的识别效率。

[0062] 进一步的，在本申请提供的一种优选实施方式中，所述视频分析模块13用于分析所述视频流子片段生成场景信息，具体用于：分析视频流子片段，生成面向对象的身份特征

识别信息和对对象动作行为的描述信息。

[0063] 这里对象的身份特征识别信息可以理解为对象的脸部特征信息。可以理解的是，获取带有某一人员对象脸部的图像，通过经预训练的识别算法对其脸部特征进行识别，即可确定该人员对象的具体身份信息。例如，确定某一学生的姓名、学号等信息，或某一教师的姓名、工号等信息。

[0064] 这里对对象动作行为的描述信息可以理解为当前视频流子片段中，人员对象的具体动作类别。可以理解的是，获取带有某一人员对象躯体动作的图像，经预训练的识别算法对该图像进行识别，即可确定该人员对象的具体动作类别。例如，确定该人员对象当前动作行为为书写行为、起立行为或板书行为等。

[0065] 通过识别视频流子片段中涉及对象的具体身份信息以及行为信息，可以确定视频流子片段对应教学场景中的师生行为，从而提升了多媒体流分析摘要的准确率。

[0066] 进一步的，在本申请提供的一种优选实施方式中，所述视频分析模块13用于分析所述视频流子片段生成第一文本信息，具体用于：分析视频流子片段，生成对象动作行为指向的第一文本信息。

[0067] 这里的对象动作行为指向的第一文本信息，可以理解为与人员对象的具体行为具有一定关联度的文本信息。可以理解的是，根据视频流子片段生成的第一文本信息中包括课堂教学场景下的文字、图片信息对应的具体文本种类或描述内容。例如，课堂中的PPT展示页、教室背景相关的黑板报信息。在课堂教学场景中，所述课堂中的PPT展示页中的文字信息，即为与人员对象动作行为相关的文本信息。但是，黑板报信息为视频流子片段中的背景信息，与当前所处场景下的人员动作行为无关，则不属于第一文本信息。

[0068] 通过针对性分析视频流子片段中与人员对象动作相关的第一文本信息，可以减少相应功能模块的数据处理量，同时增加了相应功能模块识别的识别准确度，从而有效提升了第一文本信息的分析效率。

[0069] 进一步的，在本申请提供的一种优选实施方式中，所述视频分析模块13用于分析视频流子片段，生成对象动作行为指向的第一文本信息，具体用于：分析视频流子片段，获取持续预设时长的图像；使用OCR对所述图像进行识别，生成第一文本信息。

[0070] 可以理解的是，与对象动作行为指向第一文本信息为与当前视频流子片段对应场景下，与对象行为具有一定关联度的文本信息。在课堂教学场景下，与对象行为具有一定关联度的文本信息，可以理解为教学内容相关的文本信息。例如，教师书写的板书、PPT展示页中的文字信息等。需要指出的是，在实际教学过程中，若对象动作行为指向的文本信息较为重要，则人员对象就相应文本信息展开相关行为时，会持续一定的时长。即，具有重要文本信息的图像停留时间较长。而对象动作行为指向的文本信息重要度较低或为无价值的文本信息时，则相应具有较短的时长。即，具有非重要文本信息的图像停留时间较短。因此，可以根据图像持续时长判断图像中文本的重要程度。这里的图像持续时长可以根据实际情况或者经验值预设。若某帧图像持续时长满足预设条件，即可展开后续的文本识别过程。对应的，若某帧图像持续时长不满足预设条件，说明其对应的文本重要性较低，无需展开相应的识别。这样，能够在保证第一文本识别准确度的基础上，增加第一文本的识别效率。

[0071] 具体的，对满足持续预设时长的图像记性文本信息识别时，可以通过OCR识别方式。即，利用光学字符识别技术，识别视频流子片段中满足预设条件的图像中的字符信息。

例如，通过OCR识别方式将课堂展示PPT中满足预设条件的PPT页中的相关内容识别为第一文本信息。或者，通过OCR识别方式将课堂黑板上特定标记的文字内容识别为第一文本信息。

[0072] 进一步的，在本申请提供的一种优选实施方式中，所述第一文本信息至少包括教学环节信息和知识点信息其中之一。

[0073] 这里的教学环节信息可以理解为视频流子片段中满足预设时长的图像中，记录的当前图像对应的环节。例如，PPT中当前也对应的具体标题。这里的知识点信息可以理解为视频流子片段中满足预设时长的图像中，记录的当前图像特定的标注内容。

[0074] 可以理解的是，课堂教学过程中的相关板书资料，不可避免会存在一些与教学内容无关的信息。这里将其称为无价值信息。这些无价值信息的识别，在增加计算量的同时，还会增加第一文本中无意义内容的占比。因此，通过针对性地识别视频流子片段中教学环节信息或知识点信息，便于得到更为精准的第一文本标识信息，降低了第一文本信息的冗余度，从而便于进行多媒体分析摘要的确定。

[0075] 进一步的，在本申请提供的一种优选实施方式中，所述第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。

[0076] 可以理解的是，第二文本信息对应为根据音频流子片段生成的文本信息。在课堂教学场景中，所述音频流子片段为记录有课堂教学过程中与声音信息相关的文件。在实际应用中，将音频流子片段识别为第二文本，可以用于制作与视频流子片段同步的字幕信息。过自然语言处理模型，能够对字幕中的文字信息进行纠错，并且还能进行关键词信息、提问信息、感情描述信息的提取。这些信息均可以理解为多媒体流分析摘要的形成要素。因此，经分析得到的第二文本信息至少具体包括文本纠错信息、关键词信息、提问信息、感情描述信息其中之一。这样，能够保证多媒体流分析摘要的准确性。

[0077] 进一步的，在本申请提供的一种优选实施方式中，所述分析摘要生成模块15用于处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要，具体用于：对场景信息、第一文本信息和第二文本信息进行交叉验证，形成所述多媒体流的分析摘要。

[0078] 这里所述的交叉验证可以理解为将分析得到的场景信息、第一文本信息和第二文本信息等各类数据归总到不同的课程结构中。可以理解的是，进行场景信息、第一文本信息的分析，是根据视频流子片段完成的。进行第二文本信息的识别，是根据音频流子片段完成的。所述视频流子片段与音频流子片段均为从多媒体流片段中提取得到的。若仅根据场景信息、第一文本信息和第二文本信息中的某一信息进行多媒体流的分析摘要，无法得到具有较高准确性的多媒体流分析摘要。因此，需要综合考虑场景信息、第一文本信息和第二文本信息，并将其根据课程结构进行相应的归类汇总，从而得到准确性较高的多媒体流分析摘要。这一过程，也可理解为根据获取的多媒体视频内容，完成视频内容数据的拆解，并将拆解数据重新归类关联到相应的课堂环节。

[0079] 需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，有语句“包括一个……”限定的要素，并不排除在包括所述要素

的过程、方法、商品或者设备中还存在另外的相同要素。

[0080] 以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

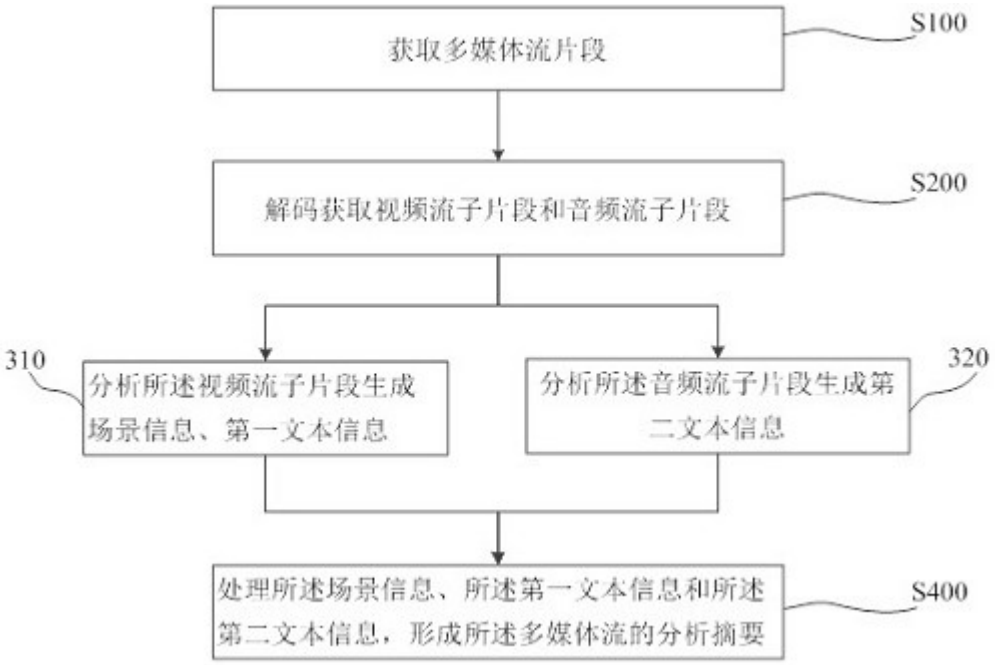

获取多媒体流片段
S100
解码获取视频流子片段和音频流子片段
S200
310
分析所述视频流子片段生成场景信息、第一文本信息
分析所述音频流子片段生成第二文本信息
320
处理所述场景信息、所述第一文本信息和所述第二文本信息，形成所述多媒体流的分析摘要
S400

图1

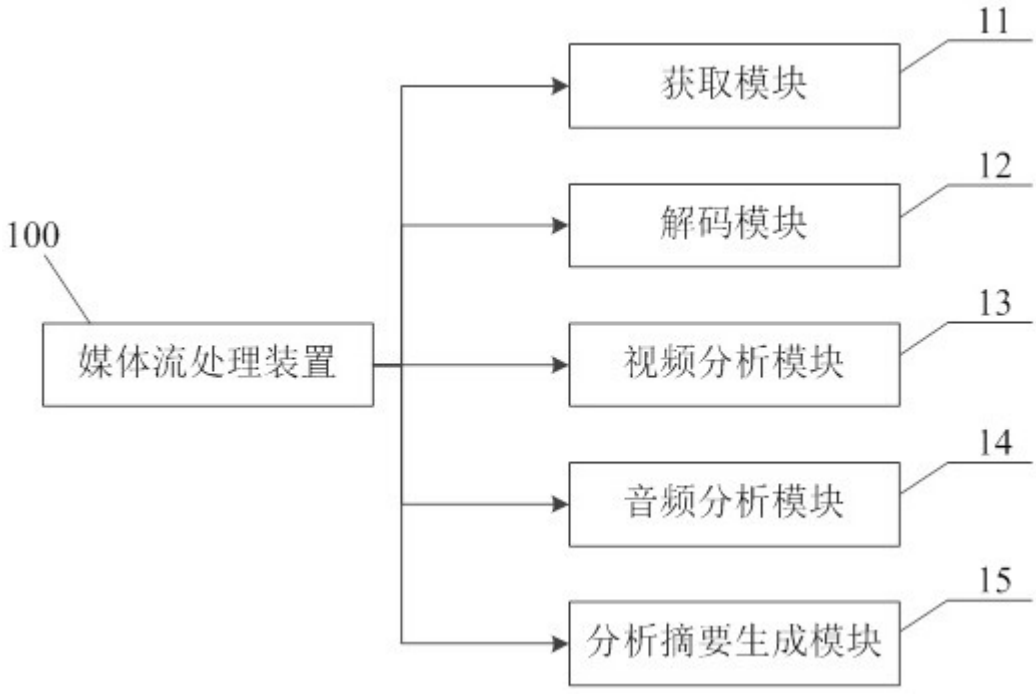

100
媒体流处理装置
获取模块
11
解码模块
12
视频分析模块
13
音频分析模块
14
分析摘要生成模块
15

图2