(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 104866596 B
(45)授权公告日 2018.09.14

(21)申请号 201510290451.8

(22)申请日 2015.05.29

(65)同一申请的已公布的文献号
申请公布号 CN 104866596 A

(43)申请公布日 2015.08.26

(73)专利权人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 李睿凡 鲁鹏 芦效峰 周延泉 李蕾 袁彩霞 刘咏彬

(74)专利代理机构 北京柏杉松知识产权代理事务所(普通合伙) 11413
代理人 马敬 项京

(51)Int.Cl.
*G06F 17/30*(2006.01)
*G06K 9/62*(2006.01)

(56)对比文件
CN 103793507 A,2014.05.14,
CN 104462489 A,2015.03.25,
Yanan Liu 等.Multimodal video classification with stacked contractive autoencoders.《Signal Processing》.2015,第120卷761-766.

审查员 邱川

权利要求书3页 说明书12页 附图6页

(54)发明名称

一种基于自动编码器的视频分类方法及装置

(57)摘要

本发明实施例提供了一种基于自动编码器的视频分类方法及装置。该方法中,获得具有三种模态数据的目标视频的每一种模态数据的低级表示内容;将每一种模态数据的低级表示内容分别输入至堆叠的自动编码器组处理,获得每一种模态数据的高级表示内容;将每两种模态数据的高级表示内容的组合结果分别输入至双模态融合器处理,获得相应两种模态数据的双模态公共表示内容;将双模态公共表示内容的组合结果输入至三模态融合器处理,获得三模态公共表示内容;将三模态公共表示内容输入至有监督分类模型处理,以获得各个视频类别的类别标签,并确定目标视频所对应的视频类别。可见,通过本方案可以结合目标视频的三种模态数据对目标视频进行分类。


CN 104866596 B

1.一种基于自动编码器的视频分类方法,其特征在于,包括:

获得目标视频的每一种模态数据的低级表示内容;其中,所述目标视频具有三种模态数据;

将所述每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理,以获得所述目标视频的每一种模态数据的高级表示内容;其中,所述堆叠的自动编码器组由至少三个自动编码器顺序相接构成,所述至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠的自动编码器组的输入内容,其余自动编码器的输入内容为前一自动编码器的隐藏层的输出内容,最后一个自动编码器的隐藏层的输出内容为所述堆叠的自动编码器组的输出内容,所述堆叠的自动编码器组的输出内容为所输入的相应模态数据的高级表示内容;

将所述目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理,以获得所述目标视频的相应两种模态数据的双模态公共表示内容;其中,所述双模态融合器为自动编码器,所述自动编码器的隐藏层的输出内容为所述双模态融合器的输出内容;

将所述目标视频的所述双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理,以获得所述目标视频的三模态公共表示内容;其中,所述三模态融合器为自动编码器,所述自动编码器的隐藏层的输出内容为所述三模态融合器的输出内容;

将所述三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理,以获得所述目标视频的各个视频类别的类别标签;其中,所述预设的有监督分类模型为基于N个样本视频所对应的三模态公共表示内容作为输入内容而相应视频样本的各个视频类别的类别标签作为输出内容所训练学习的模型;

基于所获得的所述目标视频的各个视频类别的类别标签,确定所述目标视频所对应的视频类别。

2.根据权利要求1所述的方法,其特征在于,所述有监督分类模型的构建过程包括:

获得所述N个样本视频的每一种模态数据的低级表示内容;其中,所述样本视频具有三种模态数据且对应有视频类别的类别标签;

将所述N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至所述堆叠的自动编码器组处理,以获得所述N个样本视频的每一种模态数据的高级表示内容;

将所述N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至所述双模态融合器处理,以获得所述N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容;

将所述N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至所述三模态融合器处理,以获得所述N个样本视频中每一样本视频的三模态公共表示内容;

基于所述N个样本视频中每一样本视频的三模态公共表示内容和相应的视频类别的类别标签,利用有监督学习方式,训练得到有监督分类模型。

3.根据权利要求1或2所述的方法,其特征在于,所述目标视频的图像模态数据的低级表示内容包括:色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种;

所述目标视频的音频模态数据的低级表示内容包括:MFCC特征内容,其中,MFCC为Mel

频率倒谱系数；

所述目标视频的文本模态数据的低级表示内容包括：TF-IDF特征内容，其中，所述TF-IDF为词频-逆向文档频率。

4.根据权利要求2所述的方法，其特征在于，所述有监督学习方式，包括：基于Softmax分类器的学习方式。

5.一种基于自动编码器的视频分类装置，其特征在于，包括：

低级表示内容获得模块，用于获得目标视频的每一种模态数据的低级表示内容；其中，所述目标视频具有三种模态数据；

高级表示内容获得模块，用于将所述每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得所述目标视频的每一种模态数据的高级表示内容；其中，所述堆叠的自动编码器组由至少三个自动编码器顺序相接构成，所述至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠自动编码器组的输入内容，其余自动编码器输入内容为前一自动编码器隐藏层的输出内容，最后一个自动编码器隐藏层的输出内容为所述堆叠自动编码器组的输出内容，所述堆叠自动编码器组的输出内容为所输入的相应模态数据的高级表示内容；

双模态公共表示内容获得模块，用于将所述目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得所述目标视频的相应两种模态数据的双模态公共表示内容；其中，所述双模态融合器为自动编码器，所述自动编码器隐藏层的输出内容为所述双模态融合器的输出内容；

三模态公共表示内容获得模块，用于将所述目标视频的所述双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得所述目标视频的三模态公共表示内容；其中，所述三模态融合器为自动编码器，所述自动编码器隐藏层的输出内容为所述三模态融合器的输出内容；

类别标签获得模块，用于将所述三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得所述目标视频的视频类别的类别标签；其中，所述预设的有监督分类模型为基于N个样本视频所对应的三模态公共表示内容作为输入内容而相应视频样本的视频类别的类别标签作为输出内容所训练学习的模型；

视频类别确定模块，用于基于所获得的所述目标视频的视频类别的类别标签，确定所述目标视频所对应的视频类别。

6.根据权利要求5所述的装置，其特征在于，所述有监督分类模型通过分类模型构建模块构建；

其中，所述分类模型构建模块包括：

低级表示内容获得单元，用于获得所述N个样本视频的每一种模态数据的低级表示内容；其中，所述样本视频具有三种模态数据且对应有视频类别的类别标签；

高级表示内容获得单元，用于将所述N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至所述堆叠的自动编码器组处理，以获得所述N个样本视频的每一种模态数据的高级表示内容；

双模态公共表示内容获得单元，用于将所述N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至所述双模态融合器处理，以获

得所述N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容；

三模态公共表示内容获得单元，用于将所述N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至所述三模态融合器处理，以获得所述N个样本视频中每一样本视频的三模态公共表示内容；

模型训练单元，用于基于所述N个样本视频中每一样本视频的三模态公共表示内容和相应的视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型。

7.根据权利要求5或6所述的装置，其特征在于，所述目标视频的图像模态数据的低级表示内容包括：色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种；

所述目标视频的音频模态数据的低级表示内容包括：MFCC特征内容，其中，MFCC为Mel频率倒谱系数；

所述目标视频的文本模态数据的低级表示内容包括：TF-IDF特征内容，其中，所述TF-IDF为词频-逆向文档频率。

8.根据权利要求6所述的装置，其特征在于，所述有监督学习方式，包括：基于Softmax分类器的学习方式。

# 一种基于自动编码器的视频分类方法及装置

## 技术领域

[0001] 本发明涉及视频处理技术领域，特别是涉及一种基于自动编码器的视频分类方法及装置。

## 背景技术

[0002] 为了对视频的存储、推荐、检索等进行服务，存在对大规模的视频数据进行分类处理的需求。其中，视频的类别通常包括：军事、体育、综艺、健康、生活等；并且，视频通常包括图像、音频和文本三种模态数据。

[0003] 由于每一种模态数据均为判断视频的所属类别提供了有价值的信息，因此，为了提取各种模态数据对视频分类的有价值信息，如何结合视频的三种模态数据对视频进行分类以保证分类结果的准确性，是一个亟待解决的问题。

## 发明内容

[0004] 本发明实施例的目的在于提供一种基于自动编码器的视频分类方法及装置，以结合视频的三种模态数据对视频进行分类从而保证分类结果的准确性。具体技术方案如下：

[0005] 第一方面，本发明实施例提供了一种基于自动编码器的视频分类方法，包括：

[0006] 获得目标视频的每一种模态数据的低级表示内容；其中，所述目标视频具有三种模态数据；

[0007] 将所述每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得所述目标视频的每一种模态数据的高级表示内容；其中，所述堆叠的自动编码器组由至少三个自动编码器顺序相接构成，所述至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠自动编码器组的输入内容，其余自动编码器的输入内容为前一自动编码器隐藏层的输出内容，最后一个自动编码器隐藏层的输出内容为所述堆叠自动编码器组的输出内容，所述堆叠自动编码器组的输出内容为所输入的相应模态数据的高级表示内容；

[0008] 将所述目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得所述目标视频的相应两种模态数据的双模态公共表示内容；其中，所述双模态融合器为自动编码器，所述自动编码器隐藏层的输出内容为所述双模态融合器的输出内容；

[0009] 将所述目标视频的所述双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得所述目标视频的三模态公共表示内容；其中，所述三模态融合器为自动编码器，所述自动编码器隐藏层的输出内容为所述三模态融合器的输出内容；

[0010] 将所述三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得所述目标视频的各个视频类别的类别标签；其中，所述预设的有监督分类模型为基于N个样本视频所对应的三模态公共表示内容作为输入内容而相应视频样本的各个视频类别的类别标签作为输出内容所训练学习的模型；

[0011] 基于所获得的所述目标视频的各个视频类别的类别标签，确定所述目标视频所对应的视频类别。

[0012] 可选的，所述有监督分类模型的构建过程包括：

[0013] 获得所述N个样本视频的每一种模态数据的低级表示内容；其中，所述样本视频具有三种模态数据且对应有视频类别的类别标签；

[0014] 将所述N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至所述堆叠的自动编码器组处理，以获得所述N个样本视频的每一种模态数据的高级表示内容；

[0015] 将所述N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至所述双模态融合器处理，以获得所述N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容；

[0016] 将所述N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至所述三模态融合器处理，以获得所述N个样本视频中每一样本视频的三模态公共表示内容；

[0017] 基于所述N个样本视频中每一样本视频的三模态公共表示内容和相应的视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型。

[0018] 可选的，所述目标视频的图像模态数据的低级表示内容包括：色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种；

[0019] 所述目标视频的音频模态数据的低级表示内容包括：MFCC特征内容，其中，MFCC为Mel频率倒谱系数；

[0020] 所述目标视频的文本模态数据的低级表示内容包括：TF-IDF特征内容，其中，所述TF-IDF为词频-逆向文档频率。

[0021] 可选的，所述有监督学习方式，包括：基于Softmax分类器的学习方式。

[0022] 第二方面，本发明实施例提供了一种基于自动编码器的视频分类装置，包括：

[0023] 低级表示内容获得模块，用于获得目标视频的每一种模态数据的低级表示内容；其中，所述目标视频具有三种模态数据；

[0024] 高级表示内容获得模块，用于将所述每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠自动编码器组处理，以获得所述目标视频的每一种模态数据的高级表示内容；其中，所述堆叠的自动编码器组由至少三个自动编码器顺序相接构成，所述至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠自动编码器组的输入内容，其余自动编码器的输入内容为前一自动编码器隐藏层的输出内容，最后一个自动编码器隐藏层的输出内容为所述堆叠的自动编码器组的输出内容，所述堆叠的自动编码器组的输出内容为所输入的相应模态数据的高级表示内容；

[0025] 双模态公共表示内容获得模块，用于将所述目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得所述目标视频的相应两种模态数据的双模态公共表示内容；其中，所述双模态融合器为自动编码器，所述自动编码器隐藏层的输出内容为所述双模态融合器的输出内容；

[0026] 三模态公共表示内容获得模块，用于将所述目标视频的所述双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得所述目标视频的三模态公共表

示内容;其中,所述三模态融合器为自动编码器,所述自动编码器的隐藏层的输出内容为所述三模态融合器的输出内容;

[0027] 类别标签获得模块,用于将所述三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理,以获得所述目标视频的视频类别的类别标签;其中,所述预设的有监督分类模型为基于N个样本视频所对应的三模态公共表示内容作为输入内容而相应视频样本的视频类别的类别标签作为输出内容所训练学习的模型;

[0028] 视频类别确定模块,用于基于所获得的所述目标视频的视频类别的类别标签,确定所述目标视频所对应的视频类别。

[0029] 可选的,所述有监督分类模型通过分类模型构建模块构建;

[0030] 其中,所述分类模型构建模块包括:

[0031] 低级表示内容获得单元,用于获得所述N个样本视频的每一种模态数据的低级表示内容;其中,所述样本视频具有三种模态数据且对应有视频类别的类别标签;

[0032] 高级表示内容获得单元,用于将所述N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至所述堆叠的自动编码器组处理,以获得所述N个样本视频的每一种模态数据的高级表示内容;

[0033] 双模态公共表示内容获得单元,用于将所述N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至所述双模态融合器处理,以获得所述N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容;

[0034] 三模态公共表示内容获得单元,用于将所述N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至所述三模态融合器处理,以获得所述N个样本视频中每一样本视频的三模态公共表示内容;

[0035] 模型训练单元,用于基于所述N个样本视频中每一样本视频的三模态公共表示内容和相应的视频类别的类别标签,利用有监督学习方式,训练得到有监督分类模型。

[0036] 可选的,所述目标视频的图像模态数据的低级表示内容包括:色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种;

[0037] 所述目标视频的音频模态数据的低级表示内容包括:MFCC特征内容,其中,MFCC为Mel频率倒谱系数;

[0038] 所述目标视频的文本模态数据的低级表示内容包括:TF-IDF特征内容,其中,所述TF-IDF为词频-逆向文档频率。

[0039] 可选的,所述有监督学习方式,包括:基于Softmax分类器的学习方式。

[0040] 本发明实施例中,获得具有三种模态数据的目标视频的每一种模态数据的低级表示内容;将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠自动编码器组处理,以获得目标视频的每一种模态数据的高级表示内容;将目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理,以获得目标视频的相应两种模态数据的双模态公共表示内容;将目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理,以获得目标视频的三模态公共表示内容;将三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理,以获得目标视频的各个视频类别的类别标签;基于所获得的目标视频的各个视频类别的类别标签,确定目标视频所对应的视频类别。可见,本方案结合了视频的三种模态数据对视频进行分类,从而

能够保证分类结果的准确性。

## 附图说明

[0041] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1为本发明实施例所利用的自动编码器的数据处理示意图；

[0043] 图2为本发明实施例所利用的堆叠的自动编码器组的数据处理示意图；

[0044] 图3为本发明实施例所提供的一种基于自动编码器的视频分类方法的流程图；

[0045] 图4为高级表示内容与双模态融合器、三模态融合器的关系示意图；

[0046] 图5为双模态融合器的数据处理示意图；

[0047] 图6为三模态融合器的数据处理示意图；

[0048] 图7为有监督分类模型的输入内容与输出内容的关系示意图；

[0049] 图8为本发明实施例所提供的一种基于自动编码器的视频分类方法的另一流程图；

[0050] 图9为基于三模态公共表示内容和类别标签生成有监督分类模型的示意图；

[0051] 图10为本发明实施例所提供的一种基于自动编码器的视频分类装置的结构示意图。

## 具体实施方式

[0052] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

[0053] 为清楚起见，在介绍本发明实施例所提供的一种基于自动编码器的视频分类方法及装置之前，首先对本发明实施例所需利用的现有的自动编码器和堆叠的自动编码器组进行介绍。

[0054] (1)自动编码器(Autoencoder，AE)：

[0055] 自动编码器是一种特殊的前向神经网络，其包含三层结构：输入层、隐藏层和输出层。自动编码器的特殊之处在于其输出层是输入层的重构数据，如图1，可以将输入层和隐藏层看成一个编码器，而将隐藏层和输出层看成一个解码器，那么，这种神经网络的作用就是通过对隐藏层的特定约束进行有效的编码。输入层和隐藏层所构成的编码器是一个特定的映射函数f，将输入层数据$x \in \mathbb{R}^D$，映射到隐藏层码字$y \in \mathbb{R}^d$，映射函数f如下：

[0056] $y = f(x) = f(Wx+b_h)$ (1)

[0057] 隐藏层和输出层所构成解码器所实现的解码过程是将隐藏层的码字$y \in \mathbb{R}^d$通过另一个函数g映射到数据的重构空间，形成对原始输入数据的重构$\hat{x} \in \mathbb{R}^D$，映射函数g如下：

[0058] $\hat{x}=g(y)=g(W'y+b_{\hat{x}})$ （2）

[0059] 以上编码解码的表达式中，参数b为神经网络的偏置，即$b_h \in \mathbb{R}^d$和$b_{\hat{x}} \in \mathbb{R}^D$，其中的下标分标表示隐藏层和输出重构层。参数$W \in \mathbb{R}^{D\times d}$和$W' \in \mathbb{R}^{D\times d}$分别为编码和解码的权重矩阵。对于网络中隐藏层和输出层的每个神经单元，可以使用不同的激活函数。例如，sigmoid函数，具体的，sigmoid函数如下：

[0060] $$sigmoid(u)=\frac{1}{1+e^{-u}} \quad (3)$$

[0061] 一般的，可以使编码和解码的权重矩阵参数相等，这样，可以将整个自动编码器的参数记为$\Theta=\{W,b_h,b_{\hat{x}}\}$。

[0062] 自动编码器以最小化输入数据和重构的输出数据的重构误差为目标学习整个神经网络的参数。

[0063] 假定训练数据集$D=\{x_i\}_{i=1}^N$有N个训练样本，训练目标对应的最小化代价函数为

[0064] $$J(\Theta)=\sum_{x_i\in D}L(x_i,\hat{x}_i) \quad (4)$$

[0065] 其中，L为关于某个样本的重构误差。对于线性重构，代价函数L可以使用欧式距离的平方，具体如下：

[0066] $$L(x_i,\hat{x}_i)=\left\|x_i-\hat{x}_i\right\|^2 \quad (5)$$

[0067] 对于输入数据为{0,1}类型的离散数据，L可以使用交叉熵的形式，即

[0068] $$L(x_i,\hat{x}_i)=x_i\log\hat{x}_i+(1-x_i)\log(1-\hat{x}_i) \quad (6)$$

[0069] 进一步，为防止过拟合，通常对以上形式加入惩罚项得到修正后的代价函数

[0070] $$J'(\Theta)=\sum_{x_i\in D}L(x_i,\hat{x}_i)+\frac{1}{2}\lambda\left\|W\right\|^2 \quad (7)$$

[0071] 其中，参数λ为正则化系数，控制惩罚的程度。

[0072] (2)堆叠的自动编码器组(Stacked Autoencoders，SAEs)

[0073] 自动编码器可以作为一种表示学习的模型单独使用。更为重要的是，多个自动编码器可以进行堆叠，构成堆叠式的自动编码器组，从而形成更为有效的深度表示学习模型。图2给出了一个包含三个自动编码器隐藏层的堆叠的自动编码器组。这种神经网络的学习采用逐层的方式。最初，以整个网络的输入层建立第一个自动编码器，采用自动编码器的典型学习算法；学习完毕后将该自动编码器的输出层丢弃，使用学习获得的隐藏层构建第二个自动编码器；再通过网络学习和保留隐藏层的步骤得到第三个自动编码器的输入。网络学习以这样的方式直到最顶层的自动编码器学习完毕。该顶层自动编码器的隐藏层即为整个堆叠的自动编码器组的输出层。这种无监督的训练过程称为神经网络的预训练。通过无监督的逐层学习，可以得到由预训练获得的堆叠的自动编码器组，而整个网络为进一步有监督网络的学习给出了更好地初始参数。

[0074] 基于上述的自动编码器和堆叠的自动编码器组，本发明实施例提供了一种基于自

动编码器的视频分类方法及装置，以结合视频的三种模态数据来对目标视频进行分类，从而保证分类结果的准确性。

[0075] 下面首先对本发明实施例所提供的一种基于自动编码器的视频分类方法进行介绍。

[0076] 需要说明的是，本发明实施例所针对的目标视频为具有三种模态数据的视频，其中，所述的三种模态数据包括：图像模态数据、音频模态数据和文本模态数据。

[0077] 如图3所示，本发明实施例所提供的一种基于自动编码器的视频分类方法，可以包括：

[0078] S301，获得目标视频的每一种模态数据的低级表示内容；其中，该目标视频具有三种模态数据；

[0079] 在对目标视频进行分类时，首先获得该目标视频的每一种模态数据的低级表示内容，进而对所获得的每一种模态数据的低级表示内容进行后续的处理。

[0080] 其中，该目标视频的图像模态数据的低级表示内容可以包括：色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种；该目标视频的音频模态数据的低级表示内容可以包括：MFCC特征内容，其中，MFCC为Mel频率倒谱系数；该目标视频的文本模态数据的低级表示内容可以包括：TF-IDF特征内容，其中，该TF-IDF为词频-逆向文档频率。并且，对于目标视频的图像模态数据的低级表示内容具体采用哪几种由构建有监督分类模型时所基于的样本视频的图像模态数据的低级表示内容确定。

[0081] 可以理解的是，对于图像模态数据的低级表示内容而言，色彩直方图内容用以反映图像颜色的组成分布，即各种颜色出现的概率；纹理特征内容为在二维空间变化的灰度和颜色所组成的图案；边缘特征内容为图像中灰度值发生非常明显变化的像素点组成，是图像中像素点包含不连续灰度值的结果。对于音频模态数据的低级表示内容而言，MFCC特征内容为Mel频率倒谱系数(Mel Frequency Cepstrum Coefficient，MFCC)的缩写，其中，Mel频率是基于人耳听觉特性提出来的，它与频率成非线性关系，Mel频率倒谱系数(MFCC)则是利用它们之间的这种关系，计算得到的频谱特征。

[0082] 并且，上述的每一种模态数据的低级表示内容通常为多维数据，不同类型的低级表示内容的维度不同，举例而言：色彩直方图内容、纹理特征内容和边缘特征内容的维度通常不同。并且，每一种模态数据的低级表示内容均可以采用现有技术计算得到，在此不做赘述。

[0083] 需要说明的是，在计算该目标视频的每一种模态数据的低级表示内容时可以采用采样数据。举例而言，对于音频模态数据而言，可以每间隔一段时间采集一段音频数据，间隔时间的具体时长和所采集的每段音频数据的具体时长可以根据实际情况设定；对于图像模态数据而言，可以每间隔一段时间采集至少一帧图片；而对于文本模态数据而言，可以每隔一段时间采集一定量的文字。

[0084] S302，将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得该目标视频的每一种模态数据的高级表示内容；

[0085] 在获得每一种模态数据的低级表示内容后，可以进一步将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得该目标视频的每一种模态数据的高级表示内容。

[0086] 其中，该堆叠的自动编码器组由至少三个自动编码器顺序相接构成，该至少三个自动编码器中的第一个自动编码器的输入内容为该堆叠自动编码器组的输入内容，其余自动编码器的输入内容为前一自动编码器隐藏层的输出内容，最后一个自动编码器隐藏层的输出内容为该堆叠的自动编码器组的输出内容，该堆叠的自动编码器组的输出内容为所输入的相应模态数据的高级表示内容。堆叠的自动编码器组的数据处理方式可以参见图2所示。需要说明的是，无论堆叠的自动编码器组由几个自动编码器构成，其数据处理方式均可参见图2；另外，在实际应用中，堆叠的自动编码器组的数量可以自行设定，在此不做限定。

[0087] 需要强调的是，堆叠的自编码器组的输入内容维度可以根据每一种模态数据的低级表示内容的维度来设定、下述的双模态融合器的输入内容维度可以根据每两种模态数据的高级表示内容的组合结果设定，而下述的三模态融合器的输入内容维度可以根据双模态公共表示内容的组合结果的维度设定。

[0088] S303，将该目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得该目标视频的相应两种模态数据的双模态公共表示内容；

[0089] 其中，在获得每一种模态数据的高级表示内容后，为了有效结合不同模态数据，可以进一步将该目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得该目标视频的相应两种模态数据的双模态公共表示内容。其中，该双模态融合器为自动编码器，该自动编码器的隐藏层的输出内容为该双模态融合器的输出内容。

[0090] 可以理解的是，由于目标视频存在三种模态数据，且每一种模态数据对应各自的一个低级表示内容，每一种模态数据各自对应一个高级表示内容，因此，该目标视频的每两种模态数据的高级表示内容的组合结果存在三种，如图4所示；并且，双模态融合器的数据处理方式可以参见图5所示，举例而言，图5中的每一个1可以代表文本模态数据的高级表示内容中的一维，每一个2可以代表音频模态数据的高级表示内容中的一维；另外，在组合该目标视频的每两种模态数据的高级表示内容时的组合方式在此不做赘述限定，举例而言：如图5所示的维度上的简单叠加。

[0091] S304，将该目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得该目标视频的三模态公共表示内容；

[0092] 在获得该目标视频的相应两种模态数据的双模态公共表示内容后，为了有效结合三种模态数据，可以将该目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得该目标视频的三模态公共表示内容。其中，该三模态融合器为自动编码器，该自动编码器的隐藏层的输出内容为该三模态融合器的输出内容。

[0093] 其中，三模态融合器的数据处理方式可以参见图6所示，举例而言，图6中的每个1、2、3分别可以表示一种双模态公共表示内容中的一维。

[0094] S305，将该三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得该目标视频的各个视频类别的类别标签；

[0095] 其中，预先训练得到有监督分类模型，在获得该目标视频的三模态公共表示内容后，可以将该三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得该目标视频的各个视频类别的类别标签。其中，预设的有监督分类模型为基于N个样本视

频所对应的三模态公共表示内容作为输入内容而相应视频样本的各个视频类别的类别标签作为输出内容所训练学习的模型。为了清楚起见，后续介绍有监督分类模型的构建过程。

[0096] 其中，该视频类别包括多种，具体包括哪几种视频类别是基于预先构建的有监督分类模型所确定的，并且，类别标签的表示形式可以为0和1，其中，0表示不属于本视频类别，1表示属于本视频类别。如图7所示，对于有监督分类模型而言，输出内容为三模态公共表示内容，每一组123表示三模态公共表示内容的一维，并且，输出内容为各个视频类别的类别标签，类别标签为0或1。

[0097] S306，基于所获得的该目标视频的各个视频类别的类别标签，确定该目标视频所对应的视频类别。

[0098] 由于有监督分类模型的输出结果为各个视频类别的类别标签，即每个视频类别均对应各自的类别标签，将表明属于本视频类别的类别标签所对应的视频类别确定为该目标视频所对应的视频类别。可以理解的是，对于目标视频而言，其所属的视频类别为一种，举例而言，所获得的该目标视频的各个视频类别的类别标签中存在一个值为1的类别标签，其余类别标签的值为0，那么，该目标视频所属的视频类别为：值为1的类别标签所对应的视频类别。

[0099] 本发明实施例中，获得具有三种模态数据的目标视频的每一种模态数据的低级表示内容；将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得该目标视频的每一种模态数据的高级表示内容；将目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得目标视频的相应两种模态数据的双模态公共表示内容；将目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得目标视频的三模态公共表示内容；将三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得目标视频的各个视频类别的类别标签；基于所获得的目标视频的各个视频类别的类别标签，确定目标视频所对应的视频类别。可见，本方案结合了视频的三种模态数据来对视频进行分类，从而能够保证分类结果的准确性。

[0100] 为了清楚起见，下面介绍预设的有监督分类模型的构建过程。

[0101] 具体的，如图8所示，该有监督分类模型的构建过程可以包括：

[0102] S801，获得该N个样本视频的每一种模态数据的低级表示内容；其中，该样本视频具有三种模态数据且对应有各个视频类别的类别标签；

[0103] 其中，所选取的N个样本视频为具有三种模态数据的视频，各个视频类别的类别标签是明确的，即样本视频所属的视频类别是已知的，并且，样本视频所属的视频类别为一种，举例而言：样本视频的各个视频的类别标签中存在一个值为1的类别标签，其余视频类别的类别标签均为0。另外，N的具体数据值在此不做限定，但是，可以理解的是，通常情况下，样本视频的个数越多，所训练出的有监督分类模型的精准度越高。

[0104] 其中，样本视频的图像模态数据的低级表示内容可以包括：色彩直方图内容、纹理特征内容和边缘特征内容中的至少一种；样本视频的音频模态数据的低级表示内容可以包括：MFCC特征内容，其中，MFCC为Mel频率倒谱系数；样本视频的文本模态数据的低级表示内容可以包括：TF-IDF特征内容，其中，该TF-IDF为词频-逆向文档频率。

[0105] 其中，对于图像模态数据的低级表示内容而言，色彩直方图内容用以反映图像颜

色的组成分布，即各种颜色出现的概率；纹理特征内容为在二维空间变化的灰度和颜色所组成的图案；边缘特征内容为图像中灰度值发生非常明显变化的像素点组成，是图像中像素点包含不连续灰度值的结果。对于音频模态数据的低级表示内容而言，MFCC特征内容为Mel频率倒谱系数(Mel Frequency Cepstrum Coefficient，MFCC)的缩写，其中，Mel频率是基于人耳听觉特性提出来的，它与频率成非线性关系，Mel频率倒谱系数(MFCC)则是利用它们之间的这种关系，计算得到的频谱特征。

[0106] 可以理解的是，上述的每一种模态数据的低级表示内容通常为多维数据，不同类型的低级表示内容的维度不同，举例而言：色彩直方图内容、纹理特征内容和边缘特征内容的维度通常不同。并且，每一种模态数据的低级表示内容均可以采用现有技术计算得到，在此不做赘述。

[0107] 需要说明的是，在计算N个样本视频的每一种模态数据的低级表示内容时可以采用采样数据。举例而言，对于音频模态数据而言，可以每间隔一段时间采集一段音频数据，间隔时间的具体时长和所采集的每段音频数据的具体时长可以根据实际情况设定；对于图像模态数据而言，可以每间隔一段时间采集至少一帧图片；而对于文本模态数据而言，可以每隔一段时间采集一定量的文字。

[0108] S802，将该N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至该堆叠的自动编码器组处理，以获得该N个样本视频的每一种模态数据的高级表示内容；

[0109] 其中，该堆叠的自动编码器组由至少三个自动编码器顺序相接构成，该至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠的自动编码器组的输入内容，其余自动编码器的输入内容为前一自动编码器的隐藏层的输出内容，最后一个自动编码器的隐藏层的输出内容为该堆叠的自动编码器组的输出内容，该堆叠的自动编码器组的输出内容为所输入的相应模态数据的高级表示内容，图2为堆叠的自动编码器组的数据处理示意图。需要说明的是，无论堆叠的自动编码器组由几个自动编码器构成，其数据处理方式均可参见图2；另外，在实际应用中，堆叠的自动编码器组的数量可以自行设定，在此不做限定。

[0110] 需要强调的是，构建有监督分类模型的过程中的堆叠的自动编码器组与上述的对目标视频进行视频分类时所采用的堆叠的自动编码器组是相同的，即该堆叠的自动编码器组先用于在有监督分类模型的构建过程中获得N个样本视频的每一种模态数据的高级表示内容，然后用于在目标视频的视频分类过程中的获得目标视频的每一种模态数据的高级表示内容。

[0111] 并且，堆叠的自编码器组的输入内容维度可以根据每一种模态数据的低级表示内容的维度来设定、下述的双模态融合器的输入内容维度可以根据每两种模态数据的高级表示内容的组合结果设定，而下述的三模态融合器的输入内容维度可以根据双模态公共表示内容的组合结果的维度设定。

[0112] S803，将该N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至该双模态融合器处理，以获得该N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容；

[0113] 其中，该双模态融合器为自动编码器，该自动编码器隐藏层的输出内容为该双模态融合器的输出内容。并且，在通过双模态融合器的处理后，N个样本视频中的每一样本视

频均会对应有三个双模态公共表示内容。

[0114] 可以理解的是，由于样本视频存在三种模态数据，且每一种模态数据对应各自的一个低级表示内容，每一种模态数据各自对应一个高级表示内容，因此，样本视频的每两种模态数据的高级表示内容的组合结果存在三种，如图4所示；并且，双模态融合器的数据处理方式可以参见图5所示。举例而言，图5中的每一个1可以代表文本模态数据的高级表示内容中的一维，每一个2可以代表音频模态数据的高级表示内容中的一维；另外，在组合样本视频的每两种模态数据的高级表示内容时的组合方式在此不做赘述限定，举例而言：如图5所示的维度上的简单叠加。

[0115] S804，将该N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至该三模态融合器处理，以获得该N个样本视频中每一样本视频的三模态公共表示内容；

[0116] 其中，该三模态融合器为自动编码器，该自动编码器的隐藏层的输出内容为该三模态融合器的输出内容。并且，在通过三模态融合器处理后，N个样本视频中的每一样本视频均会对应有一个三模态公共表示内容。

[0117] 其中，三模态融合器的数据处理方式可以参见图6所示，举例而言，图6中的每个1、2、3分别可以表示一种双模态公共表示内容中的一维。

[0118] S805，基于该N个样本视频中每一样本视频的三模态公共表示内容和相应的各个视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型。

[0119] 如图9所示，在获得该N个样本视频中每一样本视频的三模态公共表示内容后，可以基于该N个样本视频中每一样本视频的三模态公共表示内容和相应的各个视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型；其中，该有监督分类模型的输入内容为视频的三模态公共表示内容，输出内容为该视频的各个视频类别的类别标签。

[0120] 具体的，所述有监督学习方式，可以包括：基于Softmax分类器的学习方式。其中，基于Softmax分类器的学习方式训练得到有监督分类模型的具体过程可以采用现有技术实现，在此不做赘述。并且，上述所给出的有监督学习方式的具体方式仅仅作为示例，并不应该构成对本发明实施例的限定。

[0121] 可见，通过上述的S801～S805可以完成有监督分类模型的构建。

[0122] 相应于上述方法实施例，本发明实施例还提供了一种基于自动编码器的视频分类装置，如图10所示，可以包括：

[0123] 低级表示内容获得模块1010，用于获得目标视频的每一种模态数据的低级表示内容；其中，所述目标视频具有三种模态数据；

[0124] 高级表示内容获得模块1020，用于将所述每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得所述目标视频的每一种模态数据的高级表示内容；其中，所述堆叠的自动编码器组由至少三个自动编码器顺序相接构成，所述至少三个自动编码器中的第一个自动编码器的输入内容为所述堆叠的自动编码器组的输入内容，其余自动编码器的输入内容为前一自动编码器的隐藏层的输出内容，最后一个自动编码器的隐藏层的输出内容为所述堆叠的自动编码器组的输出内容，所述堆叠的自动编码器组的输出内容为所输入的相应模态数据的高级表示内容；

[0125] 双模态公共表示内容获得模块1030，用于将所述目标视频的每两种模态数据的高

级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得所述目标视频的相应两种模态数据的双模态公共表示内容；其中，所述双模态融合器为自动编码器，所述自动编码器的隐藏层的输出内容为所述双模态融合器的输出内容；

[0126] 三模态公共表示内容获得模块1040，用于将所述目标视频的所述双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得所述目标视频的三模态公共表示内容；其中，所述三模态融合器为自动编码器，所述自动编码器的隐藏层的输出内容为所述三模态融合器的输出内容；

[0127] 类别标签获得模块1050，用于将所述三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得所述目标视频的各个视频类别的类别标签；其中，所述预设的有监督分类模型为基于N个样本视频所对应的三模态公共表示内容作为输入内容而相应视频样本的各个视频类别的类别标签作为输出内容所训练学习的模型；

[0128] 视频类别确定模块1060，用于基于所获得的所述目标视频的各个视频类别的类别标签，确定所述目标视频所对应的视频类别。

[0129] 本发明实施例中，获得具有三种模态数据的目标视频的每一种模态数据的低级表示内容；将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得该目标视频的每一种模态数据的高级表示内容；将目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得目标视频的相应两种模态数据的双模态公共表示内容；将目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得目标视频的三模态公共表示内容；将三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得目标视频的各个视频类别的类别标签；基于所获得的目标视频的各个视频类别的类别标签，确定目标视频所对应的视频类别。可见，本方案结合了视频的三种模态数据对视频进行分类，从而能够保证分类结果的准确性。

[0130] 具体的，所述有监督分类模型通过分类模型构建模块构建；其中，所述分类模型构建模块包括：

[0131] 低级表示内容获得单元，用于获得N个样本视频的每一种模态数据的低级表示内容；其中，所述样本视频具有三种模态数据且对应有各个视频类别的类别标签；

[0132] 高级表示内容获得单元，用于将所述N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至所述堆叠的自动编码器组处理，以获得所述N个样本视频的每一种模态数据的高级表示内容；

[0133] 双模态公共表示内容获得单元，用于将所述N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至所述双模态融合器处理，以获得所述N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容；

[0134] 三模态公共表示内容获得单元，用于将所述N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至所述三模态融合器处理，以获得所述N个样本视频中每一样本视频的三模态公共表示内容；

[0135] 模型训练单元，用于基于所述N个样本视频中每一样本视频的三模态公共表示内容和相应的各个视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型。

[0136] 具体的，所述目标视频的图像模态数据的低级表示内容包括：色彩直方图内容、纹

理特征内容和边缘特征内容中的至少一种；

[0137] 所述目标视频的音频模态数据的低级表示内容包括:MFCC特征内容,其中,MFCC为Mel频率倒谱系数；

[0138] 所述目标视频的文本模态数据的低级表示内容包括:TF-IDF特征内容,其中,所述TF-IDF为词频-逆向文档频率。

[0139] 具体的,所述有监督学习方式,包括:基于Softmax分类器的学习方式。

[0140] 需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,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、物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、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“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、物品或者设备中还存在另外的相同要素。

[0141] 本说明书中的各个实施例均采用相关的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于系统实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0142] 以上所述仅为本发明的较佳实施例而已,并非用于限定本发明的保护范围。凡在本发明的精神和原则之内所作的任何修改、等同替换、改进等,均包含在本发明的保护范围内。

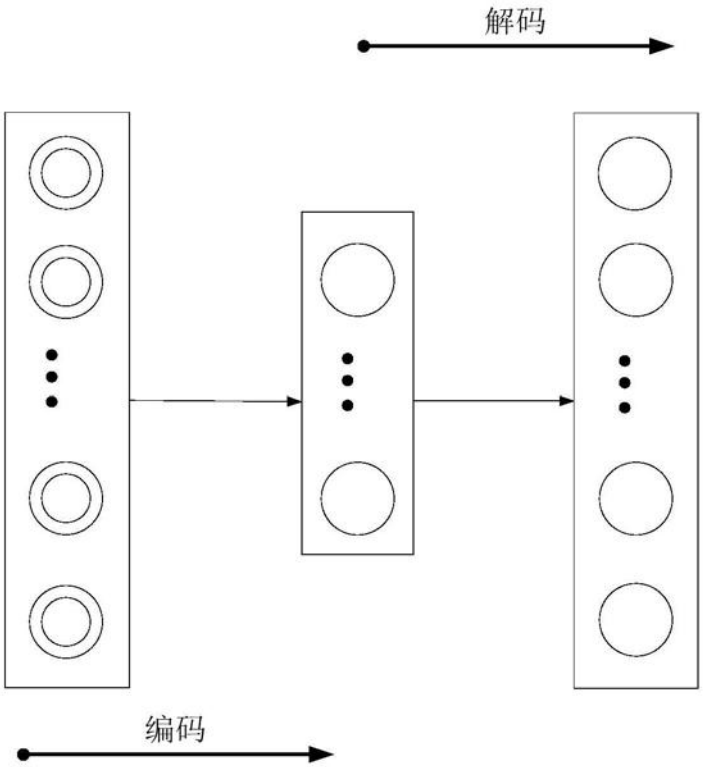


图1

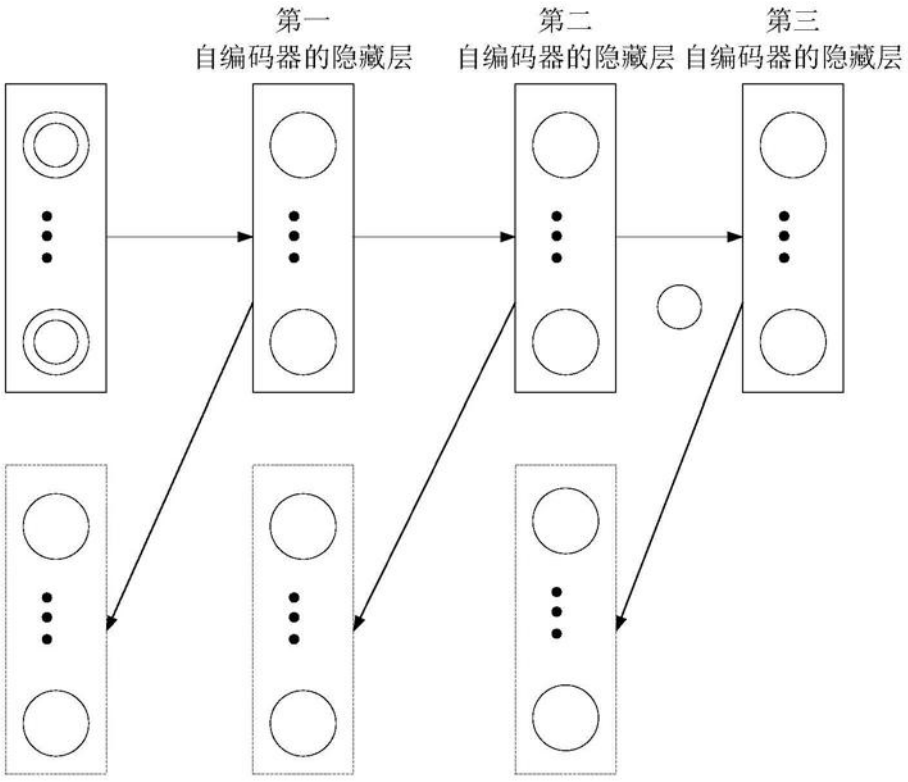


图2


获得目标视频的每一种模态数据的低级表示内容；其中，该目标视频具有三种模态数据
S301
将每一种模态数据的低级表示内容分别作为输入内容输入至预设的堆叠的自动编码器组处理，以获得该目标视频的每一种模态数据的高级表示内容
S302
将该目标视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至双模态融合器处理，以获得该目标视频的相应两种模态数据的双模态公共表示内容
S303
将该目标视频的双模态公共表示内容的组合结果作为输入内容输入至三模态融合器处理，以获得该目标视频的三模态公共表示内容
S304
将该三模态公共表示内容作为输入内容输入至预设的有监督分类模型处理，以获得该目标视频的各个视频类别的类别标签
S305
基于所获得的该目标视频的各个视频类别的类别标签，确定该目标视频所对应的视频类别
S306

图3

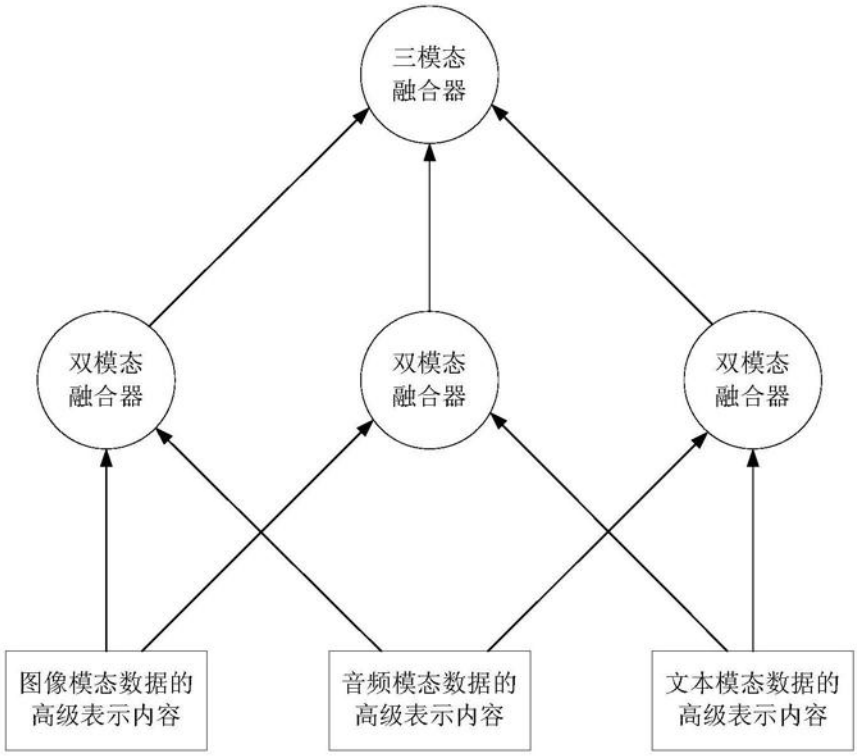
三模态
融合器
双模态
融合器
双模态
融合器
双模态
融合器
图像模态数据的
高级表示内容
音频模态数据的
高级表示内容
文本模态数据的
高级表示内容

图4

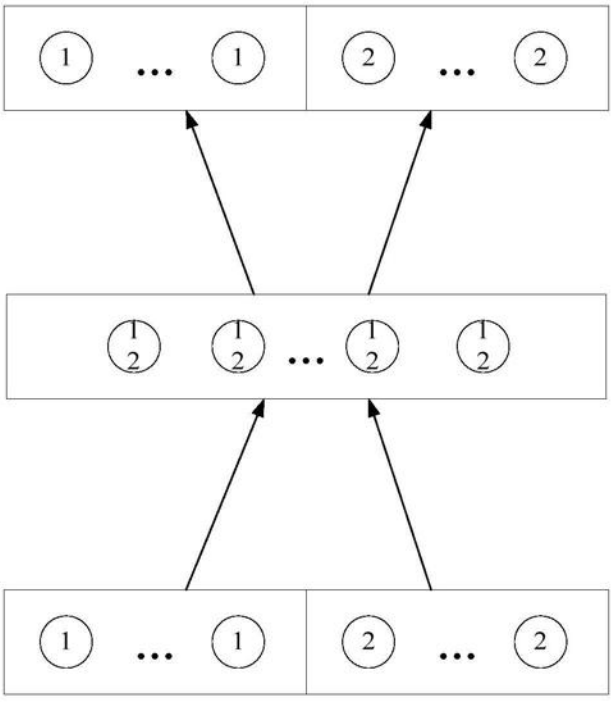
1
1
2
2
1
2
1
2
1
2
1
1
2
2

图5

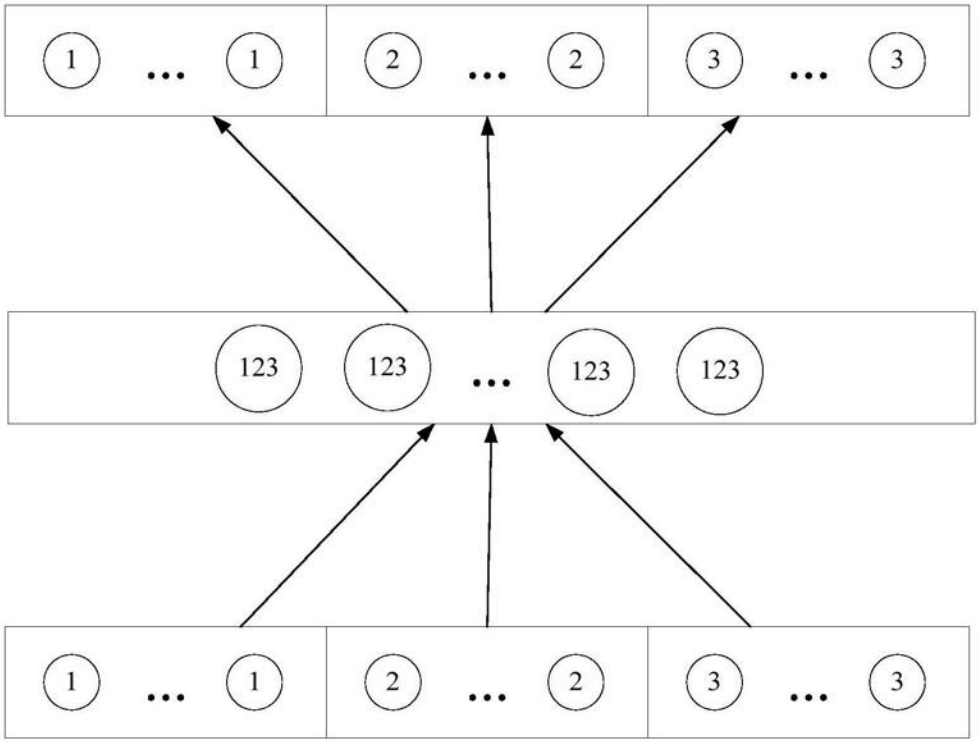


图6

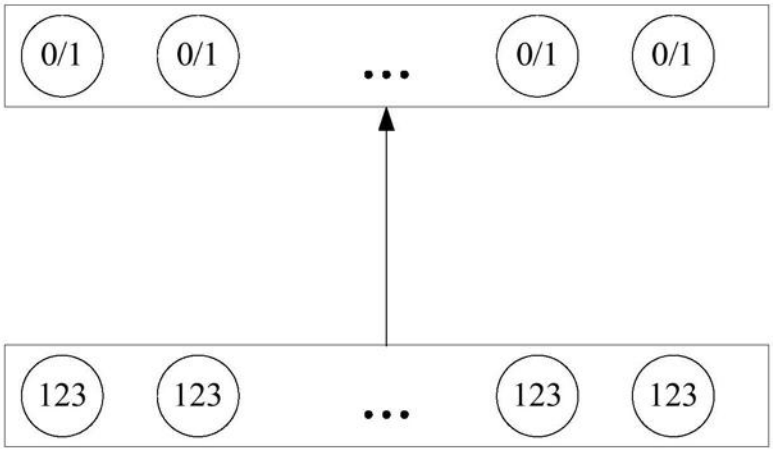


图7

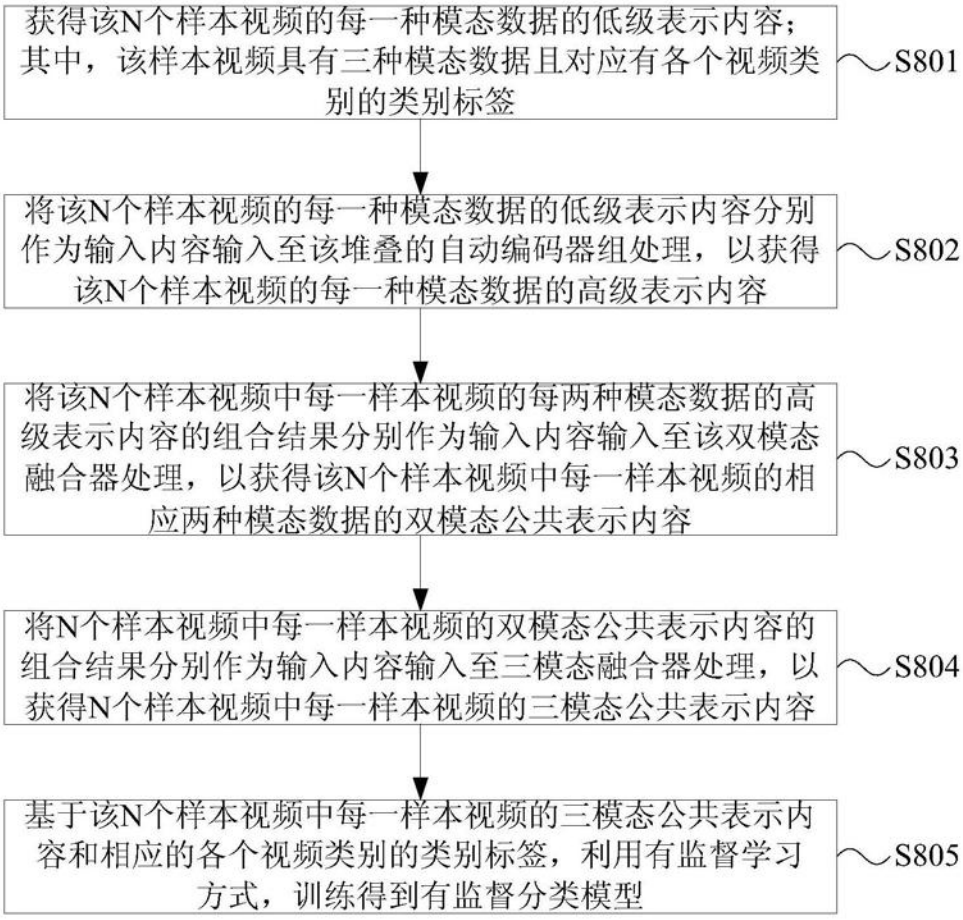
获得该N个样本视频的每一种模态数据的低级表示内容；其中，该样本视频具有三种模态数据且对应有各个视频类别的类别标签
S801
将该N个样本视频的每一种模态数据的低级表示内容分别作为输入内容输入至该堆叠的自动编码器组处理，以获得该N个样本视频的每一种模态数据的高级表示内容
S802
将该N个样本视频中每一样本视频的每两种模态数据的高级表示内容的组合结果分别作为输入内容输入至该双模态融合器处理，以获得该N个样本视频中每一样本视频的相应两种模态数据的双模态公共表示内容
S803
将N个样本视频中每一样本视频的双模态公共表示内容的组合结果分别作为输入内容输入至三模态融合器处理，以获得N个样本视频中每一样本视频的三模态公共表示内容
S804
基于该N个样本视频中每一样本视频的三模态公共表示内容和相应的各个视频类别的类别标签，利用有监督学习方式，训练得到有监督分类模型
S805

图8

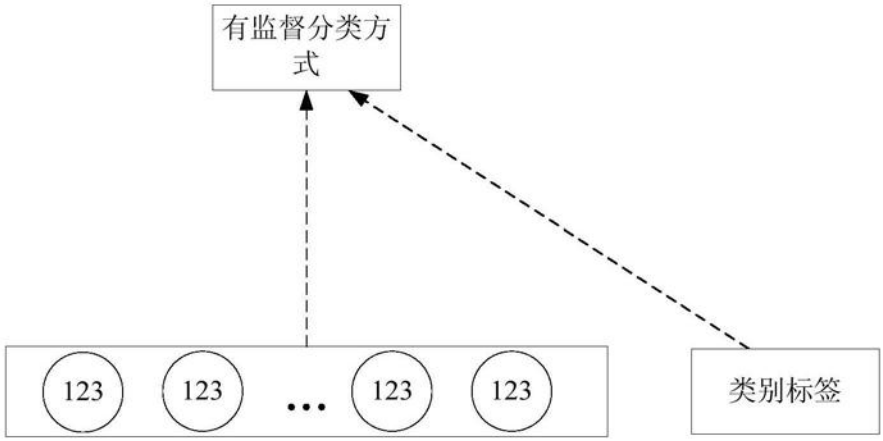
有监督分类方式
123
123
...
123
123
类别标签

图9

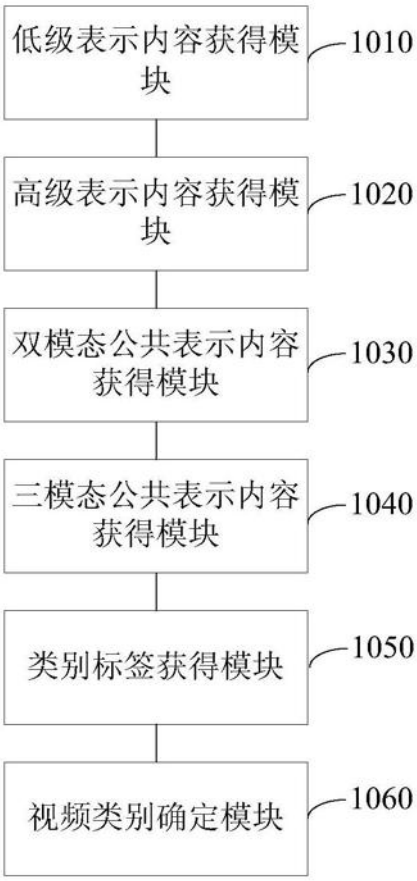
低级表示内容获得模块
1010
高级表示内容获得模块
1020
双模态公共表示内容获得模块
1030
三模态公共表示内容获得模块
1040
类别标签获得模块
1050
视频类别确定模块
1060

图10