(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 112633263 A
(43)申请公布日 2021.04.09

(21)申请号 202110253708.8
(22)申请日 2021.03.09
(71)申请人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号
(72)发明人 陶建华 连政 刘斌 孙立才
(74)专利代理机构 北京华夏泰和知识产权代理有限公司 11662
代理人 孙剑锋 刘蔓莉
(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)
G10L 25/63(2013.01)
G06N 3/04(2006.01)

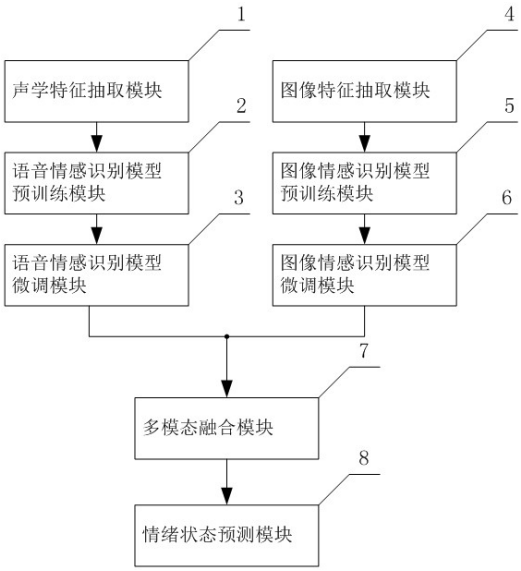
权利要求书2页 说明书7页 附图2页

(54)发明名称

海量音视频情感识别系统

(57)摘要

本申请涉及海量音视频情感识别系统，包括：声学特征抽取模块，从音频数据中抽取声学特征；语音情感识别模型预训练模块，利用海量无标注的音频数据，对语音情感识别模型进行预训练；语音情感识别模型微调模块，利用少量标注的音频数据，对语音情感识别模型进行微调；图像特征抽取模块，用于从输入的视频数据中，抽取图像特征；图像情感识别模型预训练模块，利用海量无标注的视频数据，对图像情感识别模型进行预训练；图像情感识别模型微调模块，利用少量标注的视频数据，对图像情感识别模型进行微调；多模态融合模块，将微调后的语音情感识别模型和图像情感识别模型进行融合；情绪状态预测模块，利用多模态融合后的结果，预测个体的情绪状态。



CN 112633263 A

1.海量音视频情感识别系统，其特征在于，包括：

声学特征抽取模块、语音情感识别模型预训练模块、语音情感识别模型微调模块、图像特征抽取模块、图像情感识别模型预训练模块、图像情感识别模型微调模块、多模态融合模块和情绪状态预测模块；

所述声学特征抽取模块与所述语音情感识别模型预训练模块连接，所述语音情感识别模型预训练模块与所述语音情感识别模型微调模块连接，所述图像特征抽取模块与所述图像情感识别模型预训练模块连接，所述图像情感识别模型预训练模块与所述图像情感识别模型微调模块连接，所述语音情感识别模型微调模块和所述图像情感识别模型微调模块分别与所述多模态融合模块连接，所述多模态融合模块与所述情绪状态预测模块连接；

所述声学特征抽取模块：从输入的音频数据中抽取帧级别的声学特征；

所述语音情感识别模型预训练模块：以无标注的所述帧级别的声学特征和被掩蔽的无标注的所述帧级别的声学特征为输入，预测完整的帧级别的声学特征；

所述语音情感识别模型微调模块：利用少量标注的音频数据，对所述声学特征抽取模块和所述语音情感识别模型预训练模块中的参数进行微调和优化并输出句子级别声学特征；

所述图像特征抽取模块：从输入的视频数据中抽取帧级别的图像特征；

所述图像情感识别模型预训练模块：以无标注的所述帧级别的图像特征和被掩蔽的无标注的所述帧级别的图像特征为输入，预测完整的帧级别的图像特征；

所述图像情感识别模型微调模块：利用少量标注的视频数据，对所述图像特征抽取模块和所述图像情感识别模型预训练模块中的参数进行微调和优化并输出句子级别图像特征；

所述多模态融合模块：采用特征层融合算法，将句子级别声学特征和句子级别图像特征进行拼接，作为多模态特征；

所述情绪状态预测模块：以所述多模态特征为输入，做情绪状态预测。

2.根据权利要求1所述的海量音视频情感识别系统，其特征在于，所述抽取帧级别的声学特征的方法：

首先将输入的音频按照固定的帧长和帧移进行分帧，得到帧级别的音频；

然后，从帧级别的音频中抽取韵律特征、音质特征和谱特征，作为帧级别的声学特征。

3.根据权利要求1所述的海量音视频情感识别系统，其特征在于，得到所述被掩蔽的无标注的所述帧级别的声学特征的具体方法为：

在所述无标注的所述帧级别的声学特征中随机选择10%的帧级别的声学特征，得到预掩蔽声学特征，将所述预掩蔽声学特征替换为同等维度的全零向量。

4.根据权利要求3所述的海量音视频情感识别系统，其特征在于，所述预测完整的帧级别的声学特征的具体方法：

将无标注的所述帧级别的声学特征和被掩蔽的无标注的所述帧级别的声学特征输入到循环神经网络进行训练；

计算所述循环神经网络的输出和所述帧级别的声学特征的L_2距离，作为所述语音情感识别模型预训练模块的损失函数。

5.根据权利要求1所述的海量音视频情感识别系统，其特征在于，所述语音情感识别模

型微调模块包括，声学均值池化层和声学全连接层；所述声学均值池化层和所述声学全连接层连接；

所述声学均值池化层：生成句子级别声学特征；

所述声学全连接层：预测声学句子的情绪状态。

6.根据权利要求5所述的海量音视频情感识别系统，其特征在于，采用预测声学句子的情绪状态和真实音频标注结果的交叉熵作为损失函数，对所述声学特征抽取模块和所述语音情感识别模型预训练模块中的参数进行微调和优化。

7.根据权利要求1所述的海量音视频情感识别系统，其特征在于，所述从输入的视频数据中抽取帧级别的图像特征的方法：

(1)将人脸图像作为输入，将表情标签作为输出，利用深度神经网络训练表情识别模型；

(2)将所述视频数据分解为若干图像帧；

(3)从每帧图像中识别出人脸区域位置，并进行人脸截取；

(4)将截取的人脸输入到训练好的所述表情识别模型中，抽取模型最后一层的特征作为帧级别的图像特征。

8.根据权利要求1所述的海量音视频情感识别系统，其特征在于，得到所述被掩蔽的无标注的所述帧级别的图像特征的具体方法为：

在所述无标注的所述帧级别的图像特征中随机选择10%的帧级别的图像特征，得到预掩蔽图像特征，将所述预掩蔽图像特征替换为同等维度的全零向量。

9.根据权利要求8所述的海量音视频情感识别系统，其特征在于，所述预测完整的帧级别的图像特征的具体方法：

将无标注的所述帧级别的图像特征和被掩蔽的无标注的所述帧级别的图像特征输入到循环神经网络进行训练；

计算所述循环神经网络的输出和所述帧级别的图像特征的L_2距离，作为所述图像情感识别模型预训练模块的损失函数。

10.根据权利要求1所述的海量音视频情感识别系统，其特征在于，所述图像情感识别模型微调模块包括，图像均值池化层和图像全连接层；所述图像均值池化层和所述图像全连接层连接；

所述图像均值池化层：生成句子级别图像特征；

所述图像全连接层：预测图像句子的情绪状态；

采用预测图像句子的情绪状态和真实视频标注结果的交叉熵作为损失函数，对所述图像特征抽取模块和所述图像情感识别模型预训练模块中的参数进行微调和优化。

海量音视频情感识别系统

技术领域

[0001] 本申请涉及情感识别领域，尤其涉及海量音视频情感识别系统。

背景技术

[0002] 情感识别技术不仅具有重大的科学意义，而且极具经济价值，有着广泛的应用前景。同时，其应用场景也不仅仅局限于人机交互领域，在其他诸多领域都可以发挥重要作用，例如对话生成、社会媒体分析和智能系统。

[0003] 由于情感具有一定模糊性，不同人对于相同的内容可能会有不同的感受。就像莎士比亚所说："一千个观众眼中有一千个哈姆雷特"，每个人对待任何事物都有自己的看法。为了缓解情感的模糊性，在标注过程中，我们常常需要邀请几十位专业标注人员，对情感数据进行标注，并选择标注结果的众数，作为最终的标注结果。这就导致了情感标注费时费力，很难收集大体量的标注样本。

[0004] 申请公布号CN107609572涉及多模态情感计算领域，提出了一种基于神经网络和迁移学习的多模态情感识别方法、系统，旨在解决情感数据难以获取且标注困难，使得相应识别模型不能够充分训练，造成多模态情感识别准确率不能满足需求的问题，该方法基于大规模数据训练深度神经网络并通过迁移学习获取音频特征提取器、视频特征提取器，进而对多模态情感数据进行音频特征、视频特征的提取，从而识别各语音情感类别的概率、各视频情感类别的概率，并通概率值判断最终情感类别。该方法可以有效的融合音视频两个模态，提高了多模态情感识别的准确率。

[0005] 申请公布号CN 110852215 A一种多模态情感识别方法、系统及存储介质，所述方法包括：响应所监听到的情感识别任务请求，采集音视频数据；从音视频数据中提取视频情感特征、音频情感特征和语义情感特征；将视频情感特征、音频情感特征和语义情感特征进行特征融合；依据融合情感特征进行情感特征识别。所述系统包括CPU、FPGA和存储单元；CPU能够执行前述方法步骤，FPGA能够执行前述方法中的特征提取及特征融合步骤；本发明依据视频、音频及语义情感特征融合结果进行情感特征识别，能够显著提高情感识别的准确率；将多模态情感特征提取算法同时嵌入CPU及FPGA，依据利用率选择执行设备，有助于提高算法运行速度、降低延迟。

[0006] 为了解决这一问题，本发明提供了一种基于海量音视频的情感识别系统，通过引入海量无标注的音视频数据，提升低资源情况下情感识别的性能。

发明内容

[0007] 为了解决上述技术问题或者至少部分地解决上述技术问题，本申请提供了一种海量音视频情感识别系统，包括：

声学特征抽取模块、语音情感识别模型预训练模块、语音情感识别模型微调模块、图像特征抽取模块、图像情感识别模型预训练模块、图像情感识别模型微调模块、多模态融合模块和情绪状态预测模块；所述声学特征抽取模块与所述语音情感识别模型预训练模块

连接，所述语音情感识别模型预训练模块与所述语音情感识别模型微调模块连接，所述图像特征抽取模块与所述图像情感识别模型预训练模块连接，所述图像情感识别模型预训练模块与所述图像情感识别模型微调模块连接，所述语音情感识别模型微调模块和所述图像情感识别模型微调模块分别与所述多模态融合模块连接，所述多模态融合模块与所述情绪状态预测模块连接；

所述声学特征抽取模块：从输入的音频数据中抽取帧级别的声学特征；

所述语音情感识别模型预训练模块：以无标注的所述帧级别的声学特征和被掩蔽的无标注的所述帧级别的声学特征为输入，预测完整的帧级别的声学特征；

所述语音情感识别模型微调模块：利用少量标注的音频数据，对所述声学特征抽取模块和所述语音情感识别模型预训练模块中的参数进行微调和优化并输出句子级别声学特征；

所述图像特征抽取模块：从输入的视频数据中抽取帧级别的图像特征；

所述图像情感识别模型预训练模块：以无标注的所述帧级别的图像特征和被掩蔽的无标注的所述帧级别的图像特征为输入，预测完整的帧级别的图像特征；

所述图像情感识别模型微调模块：利用少量标注的视频数据，对所述图像特征抽取模块和所述图像情感识别模型预训练模块中的参数进行微调和优化并输出句子级别图像特征；

所述多模态融合模块：采用特征层融合算法，将句子级别声学特征和句子级别图像特征进行拼接，作为多模态特征；

所述情绪状态预测模块：以所述多模态特征为输入，做情绪状态预测。

[0008] 优选的，所述抽取帧级别的声学特征的方法：

首先将输入的音频按照固定的帧长和帧移进行分帧，得到帧级别的音频；

然后，从帧级别的音频中抽取韵律特征、音质特征和谱特征，作为帧级别的声学特征。

[0009] 优选的，得到所述被掩蔽的无标注的所述帧级别的声学特征的具体方法为：

在所述无标注的所述帧级别的声学特征中随机选择10%的帧级别的声学特征，得到预掩蔽声学特征，将所述预掩蔽声学特征替换为同等维度的全零向量。

[0010] 优选的，所述预测完整的帧级别的声学特征的具体方法：

将无标注的所述帧级别的声学特征和被掩蔽的无标注的所述帧级别的声学特征输入到循环神经网络进行训练；

计算所述循环神经网络的输出和所述帧级别的声学特征的L_2距离，作为所述语音情感识别模型预训练模块的损失函数。

[0011] 优选的，所述语音情感识别模型微调模块包括，声学均值池化层和声学全连接层；所述声学均值池化层和所述声学全连接层连接；

所述声学均值池化层：生成句子级别声学特征；

所述声学全连接层：预测声学句子的情绪状态。

[0012] 优选的，采用预测声学句子的情绪状态和真实音频标注结果的交叉熵作为损失函数，对所述声学特征抽取模块和所述语音情感识别模型预训练模块中的参数进行微调和优化。

[0013] 优选的，所述从输入的视频数据中抽取帧级别的图像特征的方法：

(1)将人脸图像作为输入，将表情标签作为输出，利用深度神经网络训练表情识别模型；

(2)将所述视频数据分解为若干图像帧；

(3)从每帧图像中识别出人脸区域位置，并进行人脸截取；

(4)将截取的人脸输入到训练好的所述表情识别模型中，抽取模型最后一层的特征作为帧级别的图像特征。

[0014] 优选的，得到所述被掩蔽的无标注的所述帧级别的图像特征的具体方法为：

在所述无标注的所述帧级别的图像特征中随机选择10%的帧级别的图像特征，得到预掩蔽图像特征，将所述预掩蔽图像特征替换为同等维度的全零向量。

[0015] 优选的，所述预测完整的帧级别的图像特征的具体方法：

将无标注的所述帧级别的图像特征和被掩蔽的无标注的所述帧级别的图像特征输入到循环神经网络进行训练；

计算所述循环神经网络的输出和所述帧级别的图像特征的L_2距离，作为所述图像情感识别模型预训练模块的损失函数。

[0016] 优选的，所述图像情感识别模型微调模块包括，图像均值池化层和图像全连接层；所述图像均值池化层和所述图像全连接层连接；

所述图像均值池化层：生成句子级别图像特征；

所述图像全连接层：预测图像句子的情绪状态；

采用预测图像句子的情绪状态和真实视频标注结果的交叉熵作为损失函数，对所述图像特征抽取模块和所述图像情感识别模型预训练模块中的参数进行微调和优化。

[0017] 本申请实施例提供的上述技术方案与现有技术相比具有如下优点：

本申请实施例提供的该系统，

(1)首次提出了一种基于海量音视频的情感识别系统，通过引入海量无标注的音视频数据，提升低资源情况下情感识别的性能。

[0018] (2)采用了一种新颖的无监督预训练算法，对语音情感识别模型和图像情感识别模型进行预训练。

附图说明

[0019] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0020] 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，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0021] 图1为本申请实施例提供的海量音视频情感识别系统结构框图；

图2为本申请实施例提供的语音情感识别模型微调模块的结构框图；

图3为本申请实施例提供的图像情感识别模型微调模块的结构框图。

[0022] 图中，1—声学特征抽取模块，2—语音情感识别模型预训练模块，3—语音情感识别模型微调模块，31—声学均值池化层，32—声学全连接层，4—图像特征抽取模块，5—图

像情感识别模型预训练模块，6—图像情感识别模型微调模块，61—图像均值池化层，62—图像全连接层，7—多模态融合模块，8—情绪状态预测模块。

具体实施方式

[0023] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请的一部分实施例，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的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。

[0024] 参见图1，本申请实施例提供的海量音视频情感识别系统，包括：

声学特征抽取模块1、语音情感识别模型预训练模块2、语音情感识别模型微调模块3、图像特征抽取模块4、图像情感识别模型预训练模块5、图像情感识别模型微调模块6、多模态融合模块7和情绪状态预测模块8；

所述声学特征抽取模块1与所述语音情感识别模型预训练模块2连接，所述语音情感识别模型预训练模块2与所述语音情感识别模型微调模块3连接，所述图像特征抽取模块4与所述图像情感识别模型预训练模块5连接，所述图像情感识别模型预训练模块5与所述图像情感识别模型微调模块6连接，所述语音情感识别模型微调模块3和所述图像情感识别模型微调模块6分别与所述多模态融合模块7连接，所述多模态融合模块7与所述情绪状态预测模块8连接。

[0025] 所述声学特征抽取模块1：从输入的音频数据中抽取帧级别的声学特征，首先将输入的音频按照固定的帧长和帧移进行分帧，得到帧级别的音频；

然后，从帧级别的音频中抽取韵律特征、音质特征和谱特征，作为帧级别的声学特征。

[0026] 所述语音情感识别模型预训练模块2：以无标注的所述帧级别的声学特征和被掩蔽的无标注的所述帧级别的声学特征为输入，预测完整的帧级别的声学特征。

[0027] 在一些实施例中，得到所述被掩蔽的无标注的所述帧级别的声学特征的具体方法为：

数据库中存在N个无标注的音频数据$U=[u_1,u_2,\cdots,u_N]$，对于句子u_i，利用“声学特征抽取模块1”，抽取帧级别的声学特征$\mathbf{X}_i=[\mathbf{x}_i^1,\mathbf{x}_i^2,\cdots,\mathbf{x}_i^M]$，其中句子$u_i$总共包含$M$帧，$\mathbf{x}_i^j$表示句$u_i$子中第$j$帧的声学特征；在所述无标注的所述帧级别的声学特征$\mathbf{X}_i=[\mathbf{x}_i^1,\mathbf{x}_i^2,\cdots,\mathbf{x}_i^M]$中随机选择10%的帧级别的声学特征，得到预掩蔽声学特征，将所述预掩蔽声学特征替换为同等维度的全零向量；最终$\mathbf{X}_i$随机选择10%的帧进行掩蔽，生成的特征表示为$\widetilde{\mathbf{X}}_i$。

[0028] 在一些实施例中，所述预测完整的帧级别的声学特征的具体方法：

将$\widetilde{\mathbf{X}}_i$输入到循环神经网络进行训练，

$$\hat{\mathbf{X}}_i = RNN_A\left(\widetilde{\mathbf{X}}_i\right),$$

其中，$\hat{\mathbf{X}}_i$ 表示循环神经网络的输出；

计算所述循环神经网络的输出 $\hat{\mathbf{X}}_i$ 和所述帧级别的声学特征 $\mathbf{X}_i$ 的 L_2 距离，作为所述语音情感识别模型预训练模块2的损失函数，

$$L_{a1} = -\sum_{i=1}^{N}\left\|\hat{\mathbf{X}}_i - \mathbf{X}_i\right\|。$$

[0029] 所述语音情感识别模型微调模块3：利用少量标注的音频数据，对所述声学特征抽取模块1和所述语音情感识别模型预训练模块2中的参数进行微调和优化并输出句子级别声学特征。

[0030] 如图2所示，所述语音情感识别模型微调模块3包括，声学均值池化层31和声学全连接层32；所述声学均值池化层31和所述声学全连接层32连接；

所述声学均值池化层31：生成句子级别声学特征；

所述声学全连接层32：预测声学句子的情绪状态。

[0031] 数据库中存在K个有标注的音频数据，

$U_K = \left[(u_1, y_1), \ldots, (u_i, y_i), \ldots, (u_K, y_K)\right]$，其中 y_i 表示句子 u_i 的标签；利用“声学特征抽取模块1”，抽取帧级别的声学特征；在这一模块中，循环神经网络中的参数是经过预训练的，声学均值池化层31不包含额外的参数，只有声学全连接层32的参数是随机初始化的；

采用预测声学句子的情绪状态和真实音频标注结果的交叉熵作为损失函数，

$$L_{a2} = -\sum_{i=1}^{K} y_i \log \hat{y}_i$$

对所述声学特征抽取模块1和所述语音情感识别模型预训练模块2中的参数进行微调和优化。

[0032] 所述图像特征抽取模块4：从输入的视频数据中抽取帧级别的图像特征。

[0033] 在一些实施例中，所述从输入的视频数据中抽取帧级别的图像特征的方法：

(1)将人脸图像作为输入，将表情标签作为输出，利用深度神经网络，例如DenseNet，ResNet等网络结构，训练表情识别模型；

(2)将所述视频数据分解为若干图像帧；

(3)从每帧图像中识别出人脸区域位置，并进行人脸截取；

(4)将截取的人脸输入到训练好的所述表情识别模型中，抽取模型最后一层的特征作为帧级别的图像特征。

[0034] 所述图像情感识别模型预训练模块5：以无标注的所述帧级别的图像特征和被掩蔽的无标注的所述帧级别的图像特征为输入，预测完整的帧级别的图像特征。

[0035] 在一些实施例中，得到所述被掩蔽的无标注的所述帧级别的图像特征的具体方法

为：

数据库中存在N个无标注的图像数据$V=[v_1,v_2,...,v_N]$，对于句子$\mathbf{v}_i$，利用“图像特征抽取模块”，抽取帧级别的图像特征，$\mathbf{D}_i=\left[\mathbf{d}_i^1,\mathbf{d}_i^2,...,\mathbf{d}_i^M\right]$其中句子$\mathbf{v}_i$总共包含$M$帧，$\mathbf{d}_i^j$表示句子$\mathbf{v}_i$中第$j$帧的图像特征；在所述无标注的所述帧级别的图像特征$\mathbf{D}_i=\left[\mathbf{d}_i^1,\mathbf{d}_i^2,...,\mathbf{d}_i^M\right]$中随机选择10%的帧级别的图像特征，得到预掩蔽图像特征，将所述预掩蔽图像特征替换为同等维度的全零向量；最终$\mathbf{D}_i$随机选择10%的帧进行掩蔽，生成的特征表示为$\tilde{\mathbf{D}}_i$。

[0036] 在一些实施例中，所述预测完整的帧级别的图像特征的具体方法：

将$\tilde{\mathbf{D}}_i$输入到循环神经网络进行训练，

$$\hat{\mathbf{D}}_i=RNN_V\left(\tilde{\mathbf{D}}_i\right),$$

其中，$\hat{\mathbf{D}}_i$表示循环神经网络的输出；

计算所述循环神经网络的输出$\hat{\mathbf{D}}_i$和所述帧级别的图像特征$\mathbf{D}_i$的L_2距离，作为所述图像情感识别模型预训练模块5的损失函数，

$$L_{v1}=-\sum_{i=1}^{N}\left\|\hat{\mathbf{D}}_i-\mathbf{D}_i\right\|。$$

[0037] 所述图像情感识别模型微调模块6：利用少量标注的视频数据，对所述图像特征抽取模块4和所述图像情感识别模型预训练模块5中的参数进行微调和优化并输出句子级别图像特征。

[0038] 如图3所示，所述图像情感识别模型微调模块6包括，图像均值池化层61和图像全连接层62；所述图像均值池化层61和所述图像全连接层62连接；

所述图像均值池化层61：生成句子级别图像特征；

所述图像全连接层62：预测图像句子的情绪状态。

[0039] 数据库中存在K个有标注的视频数据，$V_K=\left[(v_1,z_1),...,(v_i,z_i),...,(v_K,z_K)\right]$，其中$z_i$表示句子$\mathbf{v}_i$的标签；利用“图像特征抽取模块4”，抽取帧级别的图像特征；在这一模块中，循环神经网络中的参数是经过预训练的，图像均值池化层61不包含额外的参数，只有图像全连接层62的参数是随机初始化的；

采用预测图像句子的情绪状态和真实视频标注结果的交叉熵作为损失函数，

$$L_{v2} = -\sum_{i=1}^{K} z_i \log \hat{z}_i$$

对所述图像特征抽取模块4和所述图像情感识别模型预训练模块5中的参数。

[0040] 所述多模态融合模块7:采用特征层融合算法,将句子级别声学特征和句子级别图像特征进行拼接,作为多模态特征。

[0041] 所述情绪状态预测模块8:以所述多模态特征为输入,做情绪状态预测。

[0042] 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。

[0043] 以上所述仅是本发明的具体实施方式,使本领域技术人员能够理解或实现本发明。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

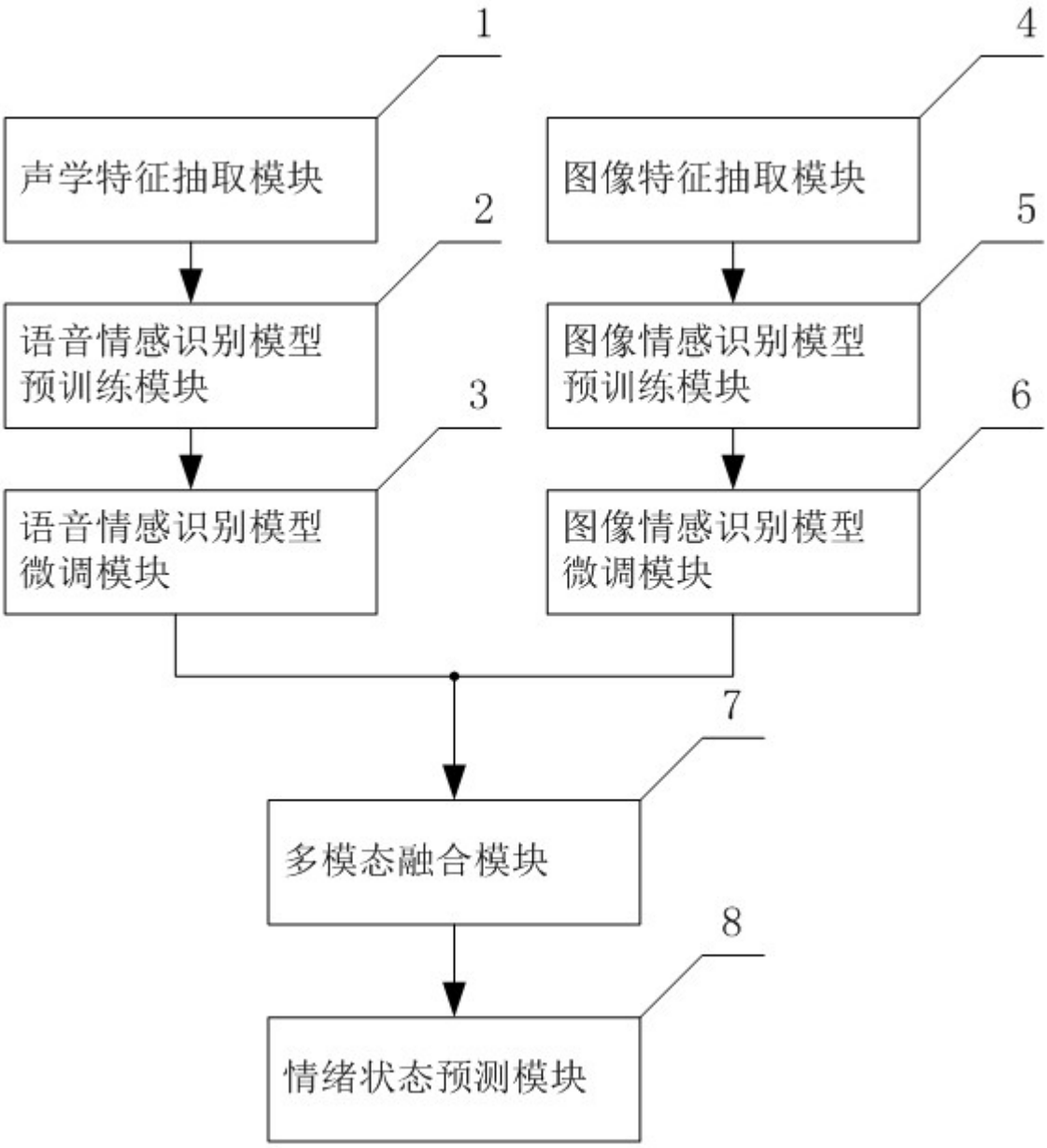

1
声学特征抽取模块
2
语音情感识别模型预训练模块
3
语音情感识别模型微调模块
4
图像特征抽取模块
5
图像情感识别模型预训练模块
6
图像情感识别模型微调模块
7
多模态融合模块
8
情绪状态预测模块

图1

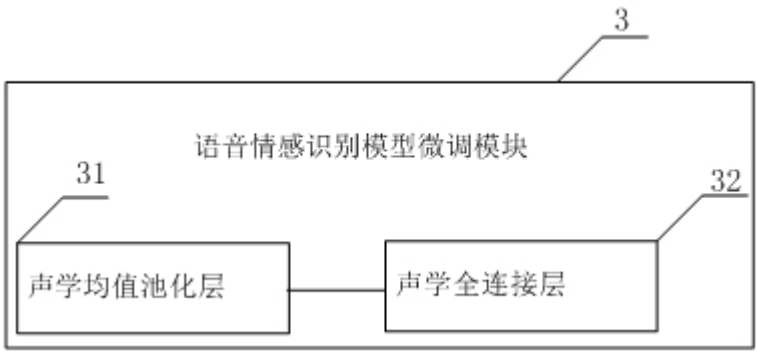

3
语音情感识别模型微调模块
31
声学均值池化层
32
声学全连接层

图2

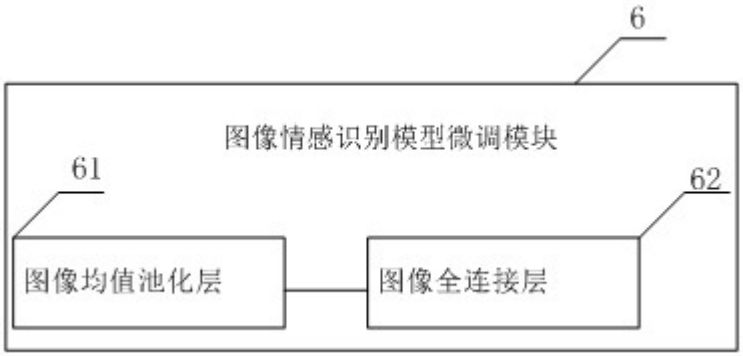


图3