

说明书

发明名称：视频行为识别方法、装置、计算机设备和存储介质

本申请要求于 2021 年 10 月 15 日提交中国专利局、申请号为 2021112027344、发明名称为“视频行为识别方法、装置、计算机设备和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，特别是涉及一种视频行为识别方法、装置、计算机设备、存储介质和计算机程序产品。

背景技术

随着计算机技术的发展，计算机视觉技术在工业、医疗、社交和导航等各领域得到了广泛的应用，通过计算机视觉，可以用计算机代替人眼对目标进行识别和测量等视觉感知处理，实现对生物视觉的模拟。视频行为识别是计算机视觉领域的重要课题之一，基于视频行为识别可以识别出给定视频出目标对象的动作行为，如吃饭、跑步、说话等各种动作行为。

目前，视频行为识别处理中，多是通过从视频中提取特征以进行行为识别，但传统的视频行为识别处理中提取的特征无法有效反映出视频中的行为信息，导致视频行为识别的准确率较低。

发明内容

根据本申请提供的各种实施例，提供一种视频行为识别方法、装置、计算机设备、存储介质和计算机程序产品。

一种视频行为识别方法，由计算机设备执行，所述方法包括：

从至少两帧目标视频图像提取视频图像特征；

将视频图像特征的空间特征进行贡献调整，得到中间图像特征；

基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的；

对融合特征进行时间特征贡献调整，得到行为识别特征；及

基于行为识别特征进行视频行为识别。

一种视频行为识别装置，所述装置包括：

视频图像特征提取模块，用于从至少两帧目标视频图像提取视频图像特征；

空间特征贡献调整模块，用于将视频图像特征的空间特征进行贡献调整，得到中间图像特征；

特征融合模块，用于基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的；

时间特征贡献调整模块，用于对融合特征进行时间特征贡献调整，得到行为识别特征；及

视频行为识别模块，用于基于行为识别特征进行视频行为识别。

一种计算机设备，包括存储器和处理器，所述存储器存储有计算机可读指令，所述处理器执行所述计算机可读指令时实现以下步骤：

从至少两帧目标视频图像提取视频图像特征；

将视频图像特征的空间特征进行贡献调整，得到中间图像特征；

基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的；

对融合特征进行时间特征贡献调整，得到行为识别特征；及

基于行为识别特征进行视频行为识别。

一种计算机可读存储介质，其上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现以下步骤：

从至少两帧目标视频图像提取视频图像特征；

将视频图像特征的空间特征进行贡献调整，得到中间图像特征；

基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的；

对融合特征进行时间特征贡献调整，得到行为识别特征；及

基于行为识别特征进行视频行为识别。

一种计算机程序产品，包括计算机可读指令，所述计算机可读指令被处理器执行时实现以下步骤：

从至少两帧目标视频图像提取视频图像特征；

将视频图像特征的空间特征进行贡献调整，得到中间图像特征；

基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的；

对融合特征进行时间特征贡献调整，得到行为识别特征；及

基于行为识别特征进行视频行为识别。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征、目的和优点将从说明书、附图以及权利要求书变得明显。

附图说明

为了更清楚地说明本申请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申请的一些实施例，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。

图 1 为一个实施例中视频行为识别方法的应用环境图；

图 2 为一个实施例中视频行为识别方法的流程示意图；

图 3 为一个实施例中对时间特征进行内聚处理的流程示意图；

图 4 为一个实施例中视频行为识别模型的结构示意图；

图 5 为一个实施例中结构参数加权融合的流程示意图；

图 6 为一个实施例中确定结构参数处理的示意图；

图 7 为一个实施例中基于先验信息进行特征融合的流程示意图；

图 8 为一个实施例中高内聚处理的流程示意图；

图 9 为一个实施例中视频行为识别装置的结构框图；

图 10 为一个实施例中计算机设备的内部结构图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供的视频行为识别方法，可以应用于如图 1 所示的应用环境中。其中，终端 102 通过网络与服务器 104 进行通信。终端 102 可以对目标对象进行拍摄，得到视频，并将获得的视频发送至服务器 104，服务器 104 从视频中提取至少两帧目标视频图像，并将至少两帧目标视频图像提取的视频图像特征的空间特征进行贡献调整，通过根据贡献调整获得的中间图像特征在时间维度的变化信息得到的先验信息，对中间图像特征的时间特征和对时间特征进行关注处理得到的内聚特征进行融合，再对得到的融合特征进行时间特征贡献调整，基于获得的行为识别特征进行视频行为识别，服务器 104 可以将得到的视频行为识别结果反馈

至终端 102。

在一些实施例中，视频行为识别方法也可以单独由服务器 104 执行，如可以由服务器 104 从数据库中获取至少两帧目标视频图像，并基于获得的至少两帧目标视频图像进行视频行为识别处理。在一些实施例中，视频行为识别方法也可以由终端 102 执行，具体可以由终端 102 拍摄到视频后，继续由终端 102 从拍摄的视频中提取至少两帧目标视频图像，并基于至少两帧目标视频图像进行视频行为识别处理。

其中，终端 102 可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑、车载设备和便携式可穿戴设备，服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在一个实施例中，如图 2 所示，提供了一种视频行为识别方法，以该方法应用于图 1 中的服务器 104 为例进行说明，包括以下步骤：

步骤 202，从至少两帧目标视频图像提取视频图像特征。

其中，目标视频图像为从需要进行行为识别处理的视频中的图像，具体可为从需要进行行为识别处理的视频中抽取的图像。例如，对于需要进行行为识别处理的视频，如为终端 102 拍摄的篮球运动视频，则目标视频图像可以为从篮球运动视频中抽取的图像。目标视频图像多于一帧，以便可以根据各帧之间的时间信息对视频进行行为识别处理。一般的，视频行为识别中，有些动作可以不需要时间信息，即不需要多帧图像之间的关联关系，只根据空间信息便能够实现行为识别，如喝水、吃饭的动作行为识别；而对于一些更细致的行为识别，则需要多帧图像之间的关联关系，即利用多帧图像之间反映的时间信息实现视频的行为识别，如对打篮球中向下拍球和向上接球的行为，需要多帧视频图像综合进行识别。在具体应用中，目标视频图像可以为从视频中连续抽取的多帧图像，如可以为连续 5 帧或 10 帧等。

视频图像特征通过对目标视频图像进行特征提取得到，用于反映目标视频图像的图像特性，视频图像特征可以为通过各种图像特征提取方式提取到的图像特征，如可以为通过人工神经网络对各帧目标视频图像进行特征提取处理提取得到的图像特征。

具体地，服务器 104 获取至少两帧目标视频图像，目标视频图像从终端 102 拍摄到的视频中提取得到，目标视频图像可以为从视频中连续抽取的多帧图像。服务器 104 从至少两帧目标视频图像中提取得到视频图像特征。具体地，服务器 104 可以对至少两帧目标视频图像分别进行图像特征提取处理，如分别输入人工神经网络中，得到各帧目标视频图像分别对应的视频图像特征。

步骤 204，将视频图像特征的空间特征进行贡献调整，得到中间图像特征。

其中，空间特征用于反映目标视频图像的空间信息，空间信息可以包括目标视频图像中各像素点的像素值分布信息，即目标视频图像中图像本身的特性。空间特征可以表征出目标视频图像包括的对象的静态特征。空间特征可以从视频图像特征中进一步提取得到，以从视频图像特征中获得反映目标视频图像中空间信息的特征。在具体实现时，可以对视频图像特征在空间维度进行特征提取，以得到视频图像特征的空间特征。贡献调整用于调整空间特征的贡献程度，空间特征的贡献程度指基于目标视频图像的特征进行视频行为识别时，空间特征对行为识别结果的影响程度。空间特征的贡献程度越大，则空间特征对视频行为识别处理的影响越大，即视频行为识别的结果越接近空间特征所反映的行为。贡献调整具体可以通过预先设定的权重参数对空间特征进行调整实现，以获得中间图像特征，中间图像特征为对视频图像特征的空间特征在视频行为识别中的贡献程度调整后得到的图像特征。

具体地，得到视频图像特征后，服务器 104 对各帧目标视频图像分别对应的视频图像特征的空间特征进行贡献调整，具体可以由服务器 104 对各视频图像特征进行空间特征提取，以提取得到各视频图像特征中的空间特征，服务器 104 基于空间权重参数对视频图像特征的空间特征分别进行贡献调整，得到中间图像特征。其中，空间权重参数可以预先设置，具体可以预先通过携带行为标签的视频图像样本训练得到。

步骤 206，基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融

合，得到融合特征；先验信息是根据中间图像特征在时间维度的变化信息得到的；内聚特征是对时间特征进行关注处理得到的。

其中，先验信息反映了目标视频图像在时间维度的先验知识，先验信息根据中间图像特征在时间维度的变化信息得到，具体可以根据中间图像特征在时间维度的相似度得到。例如，先验信息可以包括在进行特征融合时对各融合特征的权重参数，则可以将各帧目标视频图像对应的中间图像特征计算时间维度的相似度，并根据获得的相似度得到包括权重参数的先验信息。时间特征用于反映目标视频图像在视频中的时间信息，时间信息可以包括在视频中各目标视频图像之间的关联信息，即目标视频图像在视频中时间先后顺序的特性。时间特征可以表征出目标视频图像包括的对象的动态特征，从而实现对对象的动态行为识别。时间特征可以从中间图像特征中进一步提取得到，以从中间图像特征中获得反映目标视频图像中时间信息的特征。在具体实现时，可以对中间图像特征在时间维度进行特征提取，以得到中间图像特征的时间特征。时间特征对应的内聚特征是对时间特征进行关注处理得到的，关注处理指对时间特征中有利于视频行为识别的特征进行关注，以突出该特征，从而获得冗余度低、内聚性强的内聚特征，具体可以基于注意力机制的算法对中间图像特征的时间特征进行关注处理，得到时间特征对应的内聚特征。内聚特征通过对时间特征进行关注处理获得，具有高内聚性，即内聚特征的时间信息的焦点特征突出，特征冗余度低，特征有效性高，可以准确表达目标视频图像在时间维度的信息，有利于提高视频行为识别的准确率。

通过先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，以按照先验信息中的先验知识将时间特征和内聚特征进行融合，得到融合特征。融合特征通过先验信息中的先验知识将时间特征和内聚特征融合得到，可以确保融合特征中时间信息的内聚性，增强时间维度中重要特征的表达，从而能够提高视频行为识别的准确率。具体实现时，先验信息可以包括在进行特征融合时对各融合特征的权重参数，即先验信息包括时间特征和时间特征对应的内聚特征分别的权重参数，通过权重参数将时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征。

具体地，得到中间图像特征后，服务器 104 可以获取先验信息，先验信息根据中间图像特征在时间维度的变化信息得到，具体可以根据中间图像特征在时间维度的余弦相似度得到。服务器 104 基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，具体可以由服务器 104 对中间图像特征进行时间维度的特征提取，以得到中间图像特征的时间特征，并进一步确定时间特征对应的内聚特征。时间特征对应的内聚特征是通过时间特征进行关注处理得到的，具体也可以由服务器 104 基于注意力机制算法对时间特征进行关注处理，从而获得时间特征应的内聚特征。服务器 104 按照先验信息将中间图像特征的时间特征和时间特征对应的内聚特征进行融合，如服务器 104 可以按照先验信息中的权重参数，对图像特征的时间特征和时间特征对应的内聚特征进行加权融合处理，得到融合特征。

步骤 208，对融合特征进行时间特征贡献调整，得到行为识别特征。

其中，时间特征贡献调整用于调整融合特征在时间维度的贡献程度，时间特征的贡献程度指基于目标视频图像的特征进行视频行为识别时，融合特征在时间维度的特征对行为识别结果的影响程度。融合特征在时间维度的特征的贡献程度越大，则融合特征在时间维度的特征对视频行为识别处理的影响越大，即视频行为识别的结果越接近融合特征在时间维度的特征所反映的行为。时间特征贡献调整具体可以通过预先设定的权重参数对融合特征在时间维度的特征进行调整实现，以获得行为识别特征，行为识别特征可以用于视频行为识别。

具体地，获得融合特征后，服务器 104 对融合特征进行时间特征贡献调整，具体可以由服务器 104 按照时间权重参数对融合特征进行时间维度的贡献调整，以调整融合特征在时间维度的贡献程度，得到行为识别特征。其中，时间权重参数可以预先设置，具体可以预先通过携带行为标签的视频图像样本训练得到。

步骤 210，基于行为识别特征进行视频行为识别。

其中，行为识别特征为用于视频行为识别的特征，具体可以基于行为识别特征进行行为

分类，以确定目标视频图像所对应的视频行为识别结果。具体地，服务器 104 可以基于获得的行为识别特征进行视频行为识别，如可以将行为识别特征输入分类器中进行分类，根据分类结果得到视频行为识别结果，从而实现视频行为的有效识别。

上述视频行为识别方法中，将从至少两帧目标视频图像提取的视频图像特征的空间特征进行贡献调整，通过根据贡献调整获得的中间图像特征在时间维度的变化信息得到的先验信息，对中间图像特征的时间特征和对时间特征进行关注处理得到的内聚特征进行融合，再对得到的融合特征进行时间特征贡献调整，基于获得的行为识别特征进行视频行为识别。在视频行为识别处理过程中，对视频图像特征的空间特征进行贡献调整，并对融合特征进行时间特征贡献调整，可以调整行为识别特征中时间信息和空间信息的贡献程度，以增强行为识别特征的行为信息表现力；通过根据贡献调整获得的中间图像特征在时间维度的变化信息得到的先验信息，对中间图像特征的时间特征和对时间特征进行关注处理得到的内聚特征进行融合，可以对行为识别特征中时间信息进行有效聚焦，使获得的行为识别特征能够有效反映视频中的行为信息，从而提高了视频行为识别的准确率。

在一个实施例中，将视频图像特征的空间特征进行贡献调整，得到中间图像特征，包括：将视频图像特征进行空间特征提取，得到视频图像特征的空间特征；及通过结构参数中的空间结构参数对空间特征进行贡献调整，得到中间图像特征；结构参数是通过携带行为标签的视频图像样本训练得到的。

其中，空间特征提取用于从视频图像特征中提取空间特征，以对空间特征进行贡献调整。空间特征提取可以通过特征提取模块实现，如可以通过卷积神经网络模型中的卷积模块对视频图像特征进行卷积操作，以实现空间特征提取。结构参数可以包括权重参数，以对针对图像特征的各种操作处理进行加权调整。例如，对于卷积神经网络模型，结构参数可以为是卷积神经网络模型的操作空间中定义的各种操作的权重参数，具体如可以为对卷积、采样、池化等操作进行加权调整的权重参数。结构参数可以包括空间结构参数和时间结构参数，分别用于对空间维度的空间特征以及时间维度的时间特征进行贡献调整，从而调整视频图像特征中的时空信息，以增强行为识别特征的行为信息表现力，有利于提高视频行为识别的准确率。结构参数可以预先通过携带行为标签的视频图像样本训练得到，视频图像样本可以为携带行为标签的视频图像，基于视频图像样本可以训练得到结构参数，以便对各种操作处理进行有效的加权调整。

具体地，得到视频图像特征后，服务器 104 对各帧目标视频图像分别对应的视频图像特征进行空间特征提取，具体可以通过预先训练完成的视频行为识别模型对视频图像特征进行空间特征提取，如可以通过视频行为识别模型中的卷积层结构对视频图像特征进行空间特征提取，得到视频图像特征的空间特征。服务器 104 确定通过携带行为标签的视频图像样本训练得到的结构参数，并通过结构参数中的空间结构参数对空间特征进行贡献调整，如空间结构参数为权重参数时，可以通过空间结构参数所对应的权重参数对空间特征进行加权处理，以通过空间结构参数调整视频图像特征的空间特征在进行视频行为识别时对识别结果的影响程度，从而实现对空间特征的贡献调整，获得中间图像特征，中间图像特征为对视频图像特征的空间特征在视频行为识别中的贡献程度调整后得到的图像特征。

进一步地，对融合特征进行时间特征贡献调整，得到行为识别特征，包括：通过结构参数中的时间结构参数对融合特征进行贡献调整，得到行为识别特征。

其中，结构参数可以为卷积神经网络模型的操作空间中定义的各种操作的权重参数，结构参数包括对时间维度的特征进行贡献调整的时间结构参数。具体地，获得融合特征后，服务器 104 通过结构参数中的时间结构参数对融合特征进行时间特征贡献调整，得到用于视频行为处理的行为识别特征。在具体实现时，时间结构参数可以为权重参数，则服务器 104 可以通过时间结构参数所对应的权重参数对融合特征进行加权处理，以通过时间结构参数调整融合特征进行视频行为识别时，融合特征在时间维度的特征对识别结果的影响程度，从而实现对时间维度特征的贡献调整，以调整融合特征在时间维度的贡献程度，得到行为识别特征，

服务器 104 可以基于获得的行为识别特征进行视频行为识别处理，获得视频行为识别结果。

本实施例中，由通过携带行为标签的视频图像样本训练得到的结构参数中的空间结构参数和时间结构参数，分别对视频图像特征的空间特征及融合特征在相应特征维度进行贡献调整，从而根据空间结构参数和时间结构参数调整行为识别特征中时间信息和空间信息的贡献程度，实现了对时空特征的有效纠缠，使得行为识别特征的时空特征表现力强，即增强了行为识别特征的行为信息表现力，从而提高了视频行为识别的准确率。

在一个实施例中，视频行为识别方法还包括：确定待训练结构参数；通过待训练结构参数中的空间结构参数，对视频图像样本特征的空间样本特征进行贡献调整，得到中间样本特征；视频图像样本特征是从视频图像样本提取得到的；基于先验样本信息对中间样本特征的时间样本特征和时间样本特征对应的内聚样本特征进行融合，得到融合样本特征；内聚样本特征是对时间样本特征进行关注处理得到的；先验样本信息是根据中间样本特征在时间维度的变化信息得到的；通过待训练结构参数中的时间结构参数对融合样本特征进行贡献调整，得到行为识别样本特征；及基于行为识别样本特征进行视频行为识别，并根据行为识别结果和视频图像样本对应的行为标签，对待训练结构参数进行更新后继续训练直至训练结束，获得结构参数。

本实施例中，通过携带行为标签的视频图像样本进行训练，在训练结束时获得包括时间结构参数和空间结构参数的结构参数。其中，待训练结构参数可以为每次迭代训练时的初始值，通过待训练结构参数中的空间结构参数对视频图像样本特征的空间样本特征进行贡献调整，得到中间样本特征。中间样本特征为对视频图像样本特征的空间样本特征进行贡献调整后的结果，视频图像样本特征从视频图像样本提取得到，具体可以通过神经网络模型对视频图像样本进行特征提取，得到视频图像样本的视频图像样本特征。先验样本信息根据中间样本特征在时间维度的变化信息得到，具体可以为根据中间样本特征在时间维度的相似度得到；内聚样本特征通过对时间样本特征进行关注处理得到，具体可以基于注意力机制对时间样本特征进行关注处理，得到时间样本特征对应的内聚样本特征。

融合样本特征由中间样本特征的时间样本特征和时间样本特征对应的内聚样本特征按照先验样本信息进行融合得到，具体可以基于先验样本信息对中间样本特征的时间样本特征和时间样本特征对应的内聚样本特征进行加权融合，得到融合样本特征。行为识别样本特征用于视频行为识别处理，通过待训练结构参数中的时间结构参数对融合样本特征进行贡献调整得到，具体由时间结构参数对融合样本特征进行权重调整，以调整融合样本特征在时间维度的特征在视频行为识别过程中的贡献程度。行为识别结果通过基于行为识别样本特征进行视频行为识别得到，根据行为识别结果和视频图像样本对应携带的行为标签可以对待训练结构参数进行评价，根据评价结果对待训练结构参数进行更新后继续迭代训练直至训练结束，如训练次数达到预设训练次数阈值、行为识别结果满足识别精度要求、目标函数满足结束条件等，结束训练后获得训练完成的结构参数，基于训练完成的结构参数可以对视频图像特征的空间特征和融合特征分别进行贡献调整，以实现视频行为识别处理。

具体地，结构参数可以由服务器 104 训练得到，也可以由其他训练设备训练得到后移植至服务器 104 中。以服务器 104 训练结构参数为例，在训练结构参数时，服务器 104 确定待训练结构参数，待训练结构参数为当前迭代训练时的初始值，服务器 104 通过待训练结构参数中的空间结构参数，对视频图像样本特征的空间样本特征进行贡献调整，得到中间样本特征。进一步地，服务器 104 基于先验样本信息对中间样本特征的时间样本特征和时间样本特征对应的内聚样本特征进行融合，得到融合样本特征。得到融合样本特征后，服务器 104 通过待训练结构参数中的时间结构参数对融合样本特征进行贡献调整，得到行为识别样本特征，服务器 104 基于行为识别样本特征进行视频行为识别，得到行为识别结果。服务器 104 基于行为识别结果和视频图像样本对应的行为标签，对待训练结构参数进行更新，并通过更新后的待训练结构参数返回继续迭代训练直到满足训练结束条件时结束训练，获得结构参数。结构参数可以用于在进行视频行为识别处理时，对针对目标视频图像在时空维度的特征的各种

操作处理进行加权调整,从而实现目标视频图像的时空特征的有效缠绕,以增强行为识别特征的行为信息表现力,从而提高了视频行为识别的准确率。

本实施例中,通过携带行为标签的视频图像样本训练结构参数,通过训练完成的结构参数可以实现目标视频图像的时空特征的有效纠缠,能够增强行为识别特征的行为信息表现力,从而提高了视频行为识别的准确率。

在一个实施例中,视频行为识别方法通过视频行为识别模型实现,待训练结构参数是视频行为识别模型在训练中的参数。根据行为识别结果和视频图像样本对应的行为标签,对待训练结构参数进行更新后继续训练直至训练结束,获得结构参数,包括:获得视频行为识别模型输出的行为识别结果;确定行为识别结果与视频图像样本对应的行为标签之间的差异;根据差异对视频行为识别模型中的模型参数和待训练结构参数进行更新;及基于更新后的视频行为识别模型继续训练直至训练结束,并根据训练完成的视频行为识别模型得到结构参数。

本实施例中,视频行为识别方法通过视频行为识别模型实现,即通过预先训练完成的视频行为识别模型实现视频行为识别方法的步骤。视频行为识别模型可以为基于各种神经网络算法构建的人工神经网络模型,如卷积神经网络模型、深度学习网络模型、循环神经网络模型、感知机网络模型、生成对抗网络模型等。待训练结构参数是视频行为识别模型在训练中的参数,即结构参数为视频行为识别模型中对模型操作处理进行贡献调整的参数。

其中,行为识别结果为基于行为识别样本特征进行视频行为识别获得的识别结果,行为识别结果具体由视频行为识别模型输出,即将至少两帧目标视频图像输入视频行为识别模型中,以由视频行为识别模型基于目标视频图像进行视频行为识别,输出行为识别结果。行为识别结果与视频图像样本对应的行为标签之间的差异,可以通过对比行为识别结果与行为标签确定。模型参数指视频行为识别模型中各层网络结构所对应的参数。例如,对于卷积神经网络模型,模型参数可以包括但不限于包括各层卷积的卷积核参数、池化参数、上下采样参数等各种参数。通过根据行为识别结果与行为标签之间的差异对视频行为识别模型中的模型参数和待训练结构参数进行更新,以实现对视频行为识别模型中的模型参数和结构参数进行联合训练。在训练结束获得训练完成的视频行为识别模型时,根据训练完成的视频行为识别模型可以确定结构参数。

服务器 104 通过视频行为识别模型对模型参数和结构参数进行联合训练,训练完成的结构参数可以从训练完成的视频行为识别模型中确定。具体地,服务器 104 将视频图像样本输入视频行为识别模型后,由视频行为识别模型进行视频行为识别处理并输出行为识别结果。服务器 104 确定视频行为识别模型输出的行为识别结果与视频图像样本对应的行为标签之间的差异,并根据差异对视频行为识别模型的参数进行更新,具体包括对视频行为识别模型中的模型参数和待训练结构参数进行更新,得到更新后的视频行为识别模型。服务器 104 基于更新后的视频行为识别模型继续通过视频图像样本进行训练直至训练结束,如在满足训练条件时结束训练,得到训练完成的视频行为识别模型。服务器 104 可以根据训练完成的视频行为识别模型确定训练完成的结构参数,训练完成的结构参数可以对视频行为识别模型中各层网络结构的操作进行权重调整,以调整各层网络结构对视频行为识别处理的贡献程度,从而获得表现力强的特征进行视频行为识别,提高了视频行为识别的准确率。

本实施例中,通过视频行为识别模型对模型参数和结构参数进行联合训练,训练完成的结构参数可以从训练完成的视频行为识别模型中确定,通过训练完成的结构参数可以实现目标视频图像的时空特征的有效纠缠,能够增强行为识别特征的行为信息表现力,从而提高了视频行为识别的准确率。

在一个实施例中,根据行为识别结果和视频图像样本对应的行为标签,对待训练结构参数进行更新后继续训练直至训练结束,获得结构参数,包括:确定行为识别结果和视频图像样本对应的行为标签之间的行为识别损失;根据行为识别损失和前一行行为识别损失得到奖励值;及根据奖励值对待训练结构参数进行更新,通过更新后的待训练结构参数继续训练直至目标函数满足结束条件时,获得结构参数;目标函数基于训练过程中的各奖励值得到。

其中，行为识别损失用于表征行为识别结果和视频图像样本对应的行为标签之间的差异程度，行为识别损失的形式可以根据实际需要进行设置，如可以设置为交叉熵损失。前一行行为识别损失为针对前一帧视频图像样本对应确定的行为识别损失。奖励值用于对待训练结构参数进行更新，奖励值根据行为识别损失和前一行行为识别损失确定，通过奖励值可以指导待训练结构参数向满足训练要求的方向进行更新。对待训练结构参数进行更新后，通过更新后的待训练结构参数继续训练直至目标函数满足结束条件时结束训练，得到训练完成的结构参数。其中，目标函数基于训练过程中的各奖励值得到，即目标函数根据各帧视频图像样本对应的奖励值得到，具体可以根据各帧视频图像样本对应的奖励值的和构建目标函数，以根据目标函数对结构参数训练的结束进行判定，获得满足贡献调整要求的结构参数。

具体地，服务器 104 基于行为识别样本特征进行视频行为识别，得到行为识别结果后，服务器 104 确定行为识别结果和视频图像样本对应的行为标签之间的行为识别损失，具体可以通过行为识别结果与行为标签之间的交叉熵损失得到行为识别损失。服务器 104 基于获得的行为识别损失与前一帧视频图像样本对应的前一行行为识别损失得到奖励值，具体可以根据行为识别损失与前一行为识别损失之间的差异确定奖励值。例如，若行为识别损失大于前一行行为识别损失，则可以获得数值为正值的奖励值，以提供正向反馈；若行为识别损失小于前一行行为识别损失，则可以获得数值为负值的奖励值，以提供负向反馈，从而实现对待训练结构参数的更新指导。服务器 104 根据奖励值对待训练结构参数进行更新，如可以根据奖励值的正负或数值大小对待训练结构参数进行更新，得到更新后的待训练结构参数。服务器 104 通过更新后的待训练结构参数继续训练直至目标函数满足结束条件时结束训练，获得结构参数。其中，目标函数基于训练过程中的各奖励值得到，具体可以根据各帧视频图像样本对应的奖励值的和构建目标函数，通过根据目标函数对结构参数训练的结束进行判定，如在目标函数达到极值时结束训练，获得满足贡献调整要求的结构参数。

本实施例中，根据各帧视频图像样本对应的行为识别损失之间的差异得到奖励值，行为识别损失根据行为识别结果和视频图像样本对应的行为标签确定，并通过奖励值对待训练结构参数进行更新后继续进行训练，直至根据各帧视频图像样本对应的奖励值得到的目标函数满足结束条件时结束训练，得到训练完成的结构参数。通过根据各帧视频图像样本对应的行为识别损失之间的差异得到的奖励值对待训练结构参数进行更新，可以提高待训练结构参数的训练效率。

在一个实施例中，根据奖励值对待训练结构参数进行更新，包括：根据奖励值对策略梯度网络模型的模型参数进行更新；及由更新后的策略梯度网络模型对待训练结构参数进行更新。

其中，策略梯度 (Policy Gradient) 网络模型为基于策略梯度的网络模型，其输入为状态，输出为动作，策略即指在不同的状态下采取不同的动作，通过基于策略进行梯度下降，以训练策略梯度网络模型能够根据当前状态做出对应的动作，获得更高的奖励值。具体地，策略梯度网络模型的模型参数可以作为状态，而该状态下策略梯度网络模型根据输入的结构参数输出的结构参数为动作，从而策略梯度网络模型可以根据输入的结构参数和当前的模型参数预测输出下一个动作，即下一个结构参数，从而实现在训练中对结构参数的更新。

具体地，在根据奖励值对待训练结构参数进行更新时，服务器 104 根据奖励值对策略梯度网络模型的模型参数进行更新，具体基于奖励值对策略梯度网络模型中的各模型参数进行调整，以由更新后的策略梯度网络模型进行下一次的结构参数预测。对策略梯度网络模型进行更新后，服务器 104 通过更新后的策略梯度网络模型对待训练结构参数进行更新，具体可以由更新后的策略梯度网络模型基于更新后的网络状态和待训练结构参数进行结构参数预测，获得预测的结构参数，策略梯度网络模型预测的结构参数即为对待训练结构参数进行更新后的结构参数。

本实施例中，根据奖励值对策略梯度网络模型进行更新，并通过更新后的策略梯度网络模型对待训练结构参数进行更新，可以通过策略梯度方式来优化结构参数，能够确保结构参

数的训练质量，有利于提高视频行为识别处理的准确率。

在一个实施例中，由更新后的策略梯度网络模型对待训练结构参数进行更新，包括：通过更新后的策略梯度网络模型，基于更新后的模型参数和待训练结构参数进行结构参数预测，获得预测的结构参数；及根据预测的结构参数，得到对待训练结构参数进行更新后的结构参数。

其中，更新后的策略梯度网络模型，通过对策略梯度网络模型的模型参数进行更新后得到，即通过奖励值对策略梯度网络模型的模型参数进行调整更新后，得到更新后的策略梯度网络模型。

具体地，对策略梯度网络模型进行更新，得到更新后的策略梯度网络模型后，服务器以更新后的策略梯度网络模型中的模型参数作为状态，在该状态下对结构参数进行预测，具体可以基于更新后的模型参数和待训练结构参数进行结构参数预测，得到预测的结构参数。具体应用中，服务器以更新后的策略梯度网络模型的当前网络状态，利用待训练结构参数进行结构参数预测，得到预测的结构参数。服务器根据预测的结构参数进行结构参数更新，得到对待训练结构参数进行更新后的结构参数。例如，服务器可以直接将更新后的策略梯度网络模型通过结构参数预测输出的预测的结构参数，作为对待训练结构参数进行更新后的结构参数，从而实现对待训练结构参数的更新。

本实施例中，服务器通过更新后的策略梯度网络模型对待训练结构参数进行结构参数预测，并根据预测的结构参数得到对待训练结构参数进行更新后的结构参数，可以通过策略梯度方式来优化结构参数，能够确保结构参数的训练质量，有利于提高视频行为识别处理的准确率。

在一个实施例中，视频行为识别方法还包括：确定中间图像特征在时间维度的相似度；及基于相似度对初始先验信息进行修正，得到先验信息。

其中，时间维度即为各帧目标视频图像在所属视频中的先后顺序的维度，根据时间维度的时间特征，可以辅助对视频行为进行准确识别。相似度可以将表征各特征之间的距离，相似度越高，距离越近，通过中间图像特征在时间维度的相似度可以反映中间图像特征在时间维度的变化程度。初始先验信息可以为预先设定的先验信息，具体可以为预先基于样本数据训练得到的先验信息。根据相似度对初始先验信息进行修正，从而可以根据各帧目标视频图像在时间维度的变化程度，对中间图像特征的时间特征和内聚特征的融合进行加权调整，以增强融合特征的内聚性，即突出融合特征的焦点特征，减少融合特征的冗余信息。

具体地，服务器 104 在基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合前，可以根据各帧目标视频图像在时间维度的变化程度对初始先验信息进行修正，以获得对应的先验信息。服务器 104 确定中间图像特征在时间维度的相似度，具体可以对各帧目标视频图像分别对应的中间图像特征在时间维度计算余弦相似度，通过余弦相似度衡量各帧目标视频图像在时间维度的变化程度。服务器 104 根据中间图像特征在时间维度的相似度对初始先验信息进行修正，具体可以基于相似度将初始先验信息划分为正负参数，通过正负参数对初始先验信息进行修正后，将修正后的初始先验信息与初始先验信息以残差连接方式合并后得到先验信息，从而实现对先验信息的确定处理。

本实施例中，根据中间图像特征在时间维度的相似度对初始先验信息进行修正，通过反映了各帧目标视频图像在时间维度的变化程度的相似度对初始先验信息进行修正，可以有效利用各帧目标视频图像在时间维度的变化程度得到相应的先验知识，从而基于该先验知识对时间特征和内聚特征进行融合，可以对行为识别特征中时间信息进行有效聚焦，使获得的行

为识别特征能够有效反映视频中的行为信息，从而提高了视频行为识别的准确率。

在一个实施例中，初始先验信息包括第一初始先验参数和第二初始先验参数；基于相似度对初始先验信息进行修正，得到先验信息，包括：根据第一初始先验参数、第二初始先验参数及预设阈值，对相似度进行动态调整；通过动态调整后的相似度分别对第一初始先验参数和第二初始先验参数进行修正，得到第一先验参数和第二先验参数；及根据第一先验参数

和第二先验参数得到先验信息。

其中，初始先验信息包括第一初始先验参数和第二初始先验参数，第一初始先验参数和第二初始先验参数分别作为中间图像特征的时间特征以及内聚特征的融合权重参数。预设阈值可以根据实际需要进行动态设置，以根据实际需要动态修正先验信息。第一先验参数和第二先验参数分别作为中间图像特征的时间特征以及内聚特征的融合权重参数，先验信息包括第一先验参数和第二先验参数。

具体地，在对初始先验信息进行修正时，服务器 104 确定预设阈值，并根据第一初始先验参数、第二初始先验参数及预设阈值，对相似度进行动态调整。服务器 104 通过动态调整后的相似度分别对初始先验信息中的第一初始先验参数和第二初始先验参数进行修正，得到第一先验参数和第二先验参数，并根据第一先验参数和第二先验参数得到先验信息。先验信息可以对时间特征和内聚特征进行加权融合处理，以按照先验信息中的先验知识将时间特征和内聚特征进行融合，得到融合特征。融合特征通过先验信息中的先验知识将时间特征和内聚特征融合得到，可以确保融合特征中时间信息的内聚性，增强时间维度中重要特征的表达，从而能够提高视频行为识别的准确率。

本实施例中，根据初始先验信息和预设阈值对相似度进行动态调整后，基于动态调整后的相似度对分别对第一初始先验参数和第二初始先验参数进行修正，得到第一先验参数和第二先验参数，根据第一先验参数和第二先验参数得到先验信息。获得的先验信息反映了目标视频图像在时间维度的先验知识，基于该先验信息对时间特征和内聚特征进行融合，可以对行为识别特征中时间信息进行有效聚焦，使获得的行为识别特征能够有效反映视频中的行为信息，从而提高了视频行为识别的准确率。

在一个实施例中，如图 3 所示，视频行为识别方法还包括对时间特征进行内聚处理，得到对应的内聚特征的处理，具体包括：

步骤 302，确定当前基向量。

其中，当前基向量为当前对时间特征进行内聚处理的基向量，通过基向量可以实现对时间特征的内聚处理。具体地，在对时间特征进行内聚处理时，服务器 104 确定当前基向量，如可以为 $B \times C \times K$ ，其中， B 为批次处理的数据大小， C 为中间图像特征的通道数， K 为基向量的维度。

步骤 304，通过当前基向量对中间图像特征的时间特征进行特征重构，得到重构特征。

其中，由当前基向量对时间特征进行特征重构，具体可以通过当前基向量与中间图像特征的时间特征进行融合，得到重构特征。具体实现时，服务器 104 可以通过当前基向量与中间图像特征的时间特征进行矩阵相乘后进行归一化映射后，实现对时间特征的重构，得到重构特征。

步骤 306，根据重构特征和时间特征生成下一关注处理的基向量。

下一关注处理的基向量为下一次进行关注处理，即下一次对时间特征进行内聚处理时的基向量。具体地，服务器 104 根据重构特征和时间特征生成下一关注处理的基向量，如可以将重构特征和时间特征进行矩阵相乘后得到下一关注处理的基向量。下一关注处理的基向量将作为下一次进行关注处理时的基向量对相应的时间特征进行特征重构。

步骤 308，根据下一关注处理的基向量、基向量和时间特征，得到时间特征对应的内聚特征。

得到下一关注处理的基向量后，服务器 104 根据下一关注处理的基向量、基向量和时间特征，获得时间特征对应的内聚特征，从而实现对时间特征的内聚处理。具体可以将下一关注处理的基向量、基向量和时间特征进行融合后，生成时间特征对应的内聚特征。

本实施例中，通过基向量对中间图像特征的时间特征进行特征重构，根据重构特征和时间特征生成新的基向量，并根据新的基向量、旧的基向量和时间特征得到时间特征对应的内聚特征，从而对时间特征进行聚焦，以突出在时间维度的重要焦点特征，获得具有高内聚性的内聚特征，可以准确表达目标视频图像在时间维度的信息，有利于提高视频行为识别的准

确率。

在一个实施例中，根据重构特征和时间特征生成下一关注处理的基向量，包括：融合重构特征和时间特征，生成注意力特征；对注意力特征进行正则化处理，得到正则化特征；及对正则化特征进行滑动平均更新，生成下一关注处理的基向量。

其中，注意力特征通过融合重构特征和时间特征得到，通过对注意力特征依次进行正则化处理和滑动平均更新，可以确保基向量的更新更加稳定。具体地，根据重构特征和时间特征生成下一关注处理的基向量时，服务器 104 融合重构特征和时间特征得到注意力特征。服务器 104 进一步对注意力特征进行正则化处理，如可以对注意力特征进行 L2 正则化处理，得到正则化特征。服务器 104 对获得的正则化特征进行滑动平均更新，生成下一关注处理的基向量。滑动平均，或者叫做指数加权平均，可以用来估计变量的局部均值，使得变量的更新与一段时间内的历史取值有关。下一关注处理的基向量为下一次进行关注处理，即下一次对时间特征进行内聚处理时的基向量。

本实施例中，通过对融合重构特征和时间特征得到的注意力特征依次进行正则化处理和滑动平均更新，可以确保基向量的更新更加稳定，以确保内聚特征的高内聚性，可以准确表达目标视频图像在时间维度的信息，有利于提高视频行为识别的准确率。

在一个实施例中，当前基向量包括批次处理的数据大小、中间图像特征的通道数以及基向量的维度；通过当前基向量对中间图像特征的时间特征进行特征重构，得到重构特征，包括：将当前基向量与中间图像特征的时间特征，依次进行矩阵相乘及归一化映射处理，得到重构特征。

其中，批次处理的数据大小为在进行批次处理时，每个批次处理的数据量大小。例如，当前基向量可以为 $B \times C \times K$ ，其中， B 为批次处理的数据大小， C 为中间图像特征的通道数， K 为基向量的维度。具体地，服务器对中间图像特征的时间特征进行特征重构时，可以将当前基向量与中间图像特征的时间特征进行矩阵相乘，并针对矩阵相乘的结果进行归一化映射处理，实现对时间特征的重构，得到重构特征。

进一步地，根据重构特征和时间特征生成下一关注处理的基向量，包括：将重构特征和时间特征进行矩阵相乘，得到下一关注处理的基向量。

具体地，服务器将重构特征和时间特征进行矩阵相乘处理，获得下一关注处理的基向量。下一关注处理的基向量将作为下一次进行关注处理时的基向量对相应的时间特征进行特征重构。

进一步地，根据下一关注处理的基向量、基向量和时间特征，得到时间特征对应的内聚特征，包括：将下一关注处理的基向量、基向量和时间特征进行融合，得到时间特征对应的内聚特征。

具体地，服务器将下一关注处理的基向量、基向量和时间特征进行融合，从而融合下一关注处理的基向量、基向量和时间特征的有效信息，得到时间特征对应的内聚特征。

本实施例中，通过包括批次处理的数据大小、中间图像特征的通道数以及基向量的维度的基向量，对中间图像特征的时间特征进行特征重构，具体依次进行矩阵相乘及归一化映射处理，得到重构特征，并根据重构特征和时间特征进行矩阵相乘生成新的基向量，融合新的基向量、旧的基向量和时间特征得到时间特征对应的内聚特征，从而对时间特征进行聚焦，以突出在时间维度的重要焦点特征，获得具有高内聚性的内聚特征，可以准确表达目标视频图像在时间维度的信息，有利于提高视频行为识别的准确率。

在一个实施例中，基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征，包括：确定先验信息；对中间图像特征进行时间特征提取，得到中间图像特征的时间特征；及通过先验信息，对时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征。

其中，先验信息反映了目标视频图像在时间维度的先验知识，先验信息根据中间图像特征在时间维度的变化信息得到，具体可以根据中间图像特征在时间维度的相似性得到。时间

特征用于反映目标视频图像在视频中的时间信息，通过对中间图像特征进行时间特征提取，可以提取得到中间图像特征的时间特征。通过先验信息对时间特征和时间特征对应的内聚特征进行加权融合，例如在先验信息包括第一先验参数和第二先验参数时，分别通过第一先验参数和第二先验参数对时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征。

具体地，服务器 104 确定先验信息，先验信息根据中间图像特征在时间维度的变化信息得到，具体可以根据中间图像特征在时间维度的相似度得到。服务器 104 对中间图像特征进行时间特征提取，具体可以对中间图像特征中时间维度进行特征提取，以得到中间图像特征的时间特征。进一步地，服务器 104 基于先验信息对时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征，从而实现对时间特征和时间特征对应的内聚特征的加权融合，融合特征通过先验信息中的先验知识将时间特征和内聚特征融合得到，可以确保融合特征中时间信息的内聚性，增强时间维度中重要特征的表达，从而能够提高视频行为识别的准确率。

本实施例中，融合特征基于先验信息中的先验知识将时间特征和内聚特征融合得到，可以确保融合特征中时间信息的内聚性，增强时间维度中重要特征的表达，从而能够提高视频行为识别的准确率。

在一个实施例中，先验信息包括第一先验参数和第二先验参数；通过先验信息，对时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征，包括：通过第一先验参数对时间特征进行加权处理，获得加权处理后的时间特征；通过第二先验参数对时间特征对应的内聚特征进行加权处理，得到加权处理后的内聚特征；及将加权处理后的时间特征和加权处理后的内聚特征进行融合，得到融合特征。

其中，先验信息包括第一先验参数和第二先验参数，分别对应于时间特征和时间特征对应的内聚特征的加权重。具体地，服务器通过先验信息中的第一先验参数，对时间特征进行加权处理，获得加权处理后的时间特征。例如，第一先验参数可以为 k_1 ，时间特征可以为 M ，则加权处理后的时间特征可以为 $k_1 * M$ 。服务器通过先验信息中的第二先验参数，时间特征对应的内聚特征进行加权处理，获得加权处理后的内聚特征。例如，第二先验参数可以为 k_2 ，时间特征对应的内聚特征可以为 N ，则加权处理后的内聚特征可以为 $k_2 * N$ 。服务器将加权处理后的时间特征和加权处理后的内聚特征进行融合，得到融合特征，如服务器融合得到的融合特征可以为 $k_1 * M + k_2 * N$ 。

本实施例中，融合特征基于先验信息中的第一先验参数和第二先验参数将时间特征和内聚特征融合得到，可以确保融合特征中时间信息的内聚性，增强时间维度中重要特征的表达，从而能够提高视频行为识别的准确率。

在一个实施例中，在基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征之前，还包括：对中间图像特征进行标准化处理，得到标准化特征；及根据标准化特征进行非线性映射，获得映射后的中间图像特征。

其中，标准化处理可以对中间图像特征进行规范化，有利于解决梯度消失和梯度爆炸问题，能够确保网络学习速率。标准化处理可以通过批量标准化处理实现。非线性映射可以引入非线性因素，从而对中间图像特征进行去线性，有利于增强中间图像特征的灵活表达。具体地，得到中间图像特征后，服务器 104 对中间图像特征进行标准化处理，如可以通过 BN（Batch Normalization，批量标准化）层结构对中间图像特征进行标准化处理，得到标准化特征。进一步地，服务器 104 对标准化特征进行非线性映射，如可以通过激活函数对标准化特征进行非线性映射，得到映射后的中间图像特征。

进一步地，基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征，包括：基于先验信息对映射后的中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据映射后的中间图像特征在时间维度的变化信息得到的。

具体地，得到映射后的中间图像特征后，服务器 104 基于先验信息对映射后的中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征。其中，先验信息是根

据映射后的中间图像特征在时间维度的变化信息得到的，内聚特征是对映射后的中间图像特征的时间特征进行关注处理得到的。

本实施例中，在得到中间图像特征后，进一步对中间图像特征进行标准化处理和非线性映射，以增强中间图像特征的特征表达，并基于映射后的中间图像特征进行视频行为识别处理，可以进一步提高行为识别特征的行为信息表现力，从而有利于提高视频行为识别的准确率。

在一个实施例中，对中间图像特征进行标准化处理，得到标准化特征，包括：通过批量标准化层结构，对中间图像特征进行标准化处理，得到标准化特征。

其中，批量标准化层结构为 BN 层结构，可以对中间图像特征批量进行标准化处理。具体地，服务器可以通过批量标准化层结构对中间图像特征批量进行标准化处理，得到标准化特征，从而能够确保标准化的处理效率。

进一步地，根据标准化特征进行非线性映射，获得映射后的中间图像特征，包括：通过激活函数对标准化特征进行非线性映射，获得映射后的中间图像特征。

其中，激活函数用于引入非线性因素，以实现对标准化特征的非线性映射。激活函数的具体形式可以根据实际需要进行设置，如可以设置 ReLU 函数，以由服务器通过激活函数对标准化特征进行非线性映射，得到映射后的中间图像特征。

本实施例中，在得到中间图像特征后，进一步通过批量标准化层结构、激活函数对中间图像特征，依次进行标准化处理和非线性映射，以增强中间图像特征的特征表达，并提高处理效率。

本申请还提供一种应用场景，该应用场景应用上述的视频行为识别方法。具体地，该视频行为识别方法在该应用场景的应用如下：

对于视频行为识别处理，时空信息建模是视频行为识别的核心问题之一。近年来主流方法主要有基于双流网络的行为识别方法和基于 3D（3-Dimensional，三维）卷积网络的行为识别方法。前者通过平行的两个网络分别提取 RGB 和光流特征，后者通过 3D 卷积同时建模时间和空间信息。然而，大量的模型参数和算力损耗限制了其效率，基于此，后续的改进方法主要通过将三维卷积分解为二维空间卷积和一维时间卷积的方式来分别对时间和空间信息建模，进而提升模型的效率。

通过设计不同的网络结构来提取更好的时空特征，但忽略了时空线索对不同动作类的差异化影响。例如，有些动作即使没有时间信息的帮助，也很容易仅用一张图片来判别，这是因为它们在不同的场景中，具有显著的空间信息，此时可以作为具有高度可信度的动作类别进行预测。然而，时间信息对细粒度动作识别是必不可少的，例如，“拉小提琴”动作中的推弦弓和拉弦弓动作的判别，需要时间信息才可以针对推弦弓和拉弦弓动作进行准确识别。视频中通常包含丰富的时间相互关联的内容，在这样多维的信息中，仅仅对时空特征进行独立分解建模，而时空信息的相关性在不同的动作类别之间存在很大的差异，在识别过程中对时空信息的贡献不同，导致时空信息无法有效反映出视频中的行为信息。此外，视频中动作的时间边界不明确，即动作的开始时间和结束时间不明确、持续时间不确定，导致视频行为识别的准确率较低。

基于此，本实施例中通过上述的视频行为识别方法，可以采用网络结构搜索策略自适应地调整时间和空间信息的权重，根据行为识别过程中贡献的不同，挖掘时间空间信息之间的深层关联、共同学习时空的相互作用；同时设计了一个节奏调节器，根据动作节律的先验信息和时间卷积的结构参数，得到时间信息的高内聚性表达，以此来调整不同节奏的动作，从而解决相同动作却具有不同节奏造成的特征表达差异的问题，提高了视频行为识别的准确率。

具体地，视频行为识别方法包括：从至少两帧目标视频图像提取视频图像特征，具体可以将至少两帧目标视频图像输入到人工神经网络中，以由人工神经网络提取得到视频图像特征；将视频图像特征的空间特征进行贡献调整，得到中间图像特征，具体通过预先训练的结构参数对视频图像特征的空间特征进行贡献调整；基于先验信息对中间图像特征的时间特征

和时间特征对应的内聚特征进行融合，从而使用节奏调节器调整行为的节奏，得到融合特征；再对融合特征进行时间特征贡献调整，得到行为识别特征，具体可以通过结构参数对融合特征进行时间特征贡献调整；最后基于行为识别特征进行视频行为识别，得到行为识别结果。

本实施例的视频行为识别方法基于视频行为识别模型实现，如图 4 所示，为本实施例中视频行为识别模型的网络结构示意图。其中， X 为至少两帧目标视频图像提取到的视频图像特征，通过 $1 \times 3 \times 3$ 的 2D 卷积进行空间特征提取，得到空间特征，并通过结构参数中的空间结构参数 α_1 对空间特征进行贡献调整，得到中间图像特征。中间图像特征依次通过批量标准化处理和激活函数的非线性映射处理，具体可以通过 BN 层结构和 ReLU 层结构实现对中间图像特征的批量标准化处理和非线性映射处理。获得的映射后的特征 A 分别通过两个 $3 \times 1 \times 1$ 的 1D 卷积进行时间特征提取，其中一个分支为高内聚 Cohesive 的 1D 卷积处理，从而可以提取得到中间图像特征的时间特征对应的内聚特征。对于 1D 卷积进行时间特征提取的结果，通过先验信息中的权重参数 β_1 和 β_2 分别进行加权调整，并对两个分支的加权调整结果进行融合。权重参数 β_1 和 β_2 可以为基于策略梯度 Agent 网络训练得到的结构参数，通过确定特征 A 在时间维度的相似度，以对初始的权重参数 β_1 和 β_2 进行残差修正，并基于残差修正后的权重参数 β_1 和 β_2 对 1D 卷积的提取结果进行加权处理。两个 1D 卷积分支的结果进行融合后，通过结构参数中的时间结构参数 α_2 对融合特征进行时间特征贡献调整，对贡献调整后的融合特征进行下采样后得到行为识别特征，行为识别特征用于视频行为识别，得到行为识别结果。

其中，结构参数指的是操作空间中定义的诸如卷积等操作的权重参数，是网络结构搜索技术中的概念。本实施例可以通过微分方式和策略梯度方式两种结构参数更新方式来优化更新待融合的时间和空间卷积对应的结构参数，包括 α_1 和 α_2 ；而在高内聚时间卷积模块和 1D 时间卷积模块的融合中，也可以利用预先训练的结构参数 β_1 和 β_2 进行加权融合处理。如图 5 所示，融合时间和空间卷积的结构参数包括 α_1 和 α_2 ，对两个时间卷积分支进行加权融合的结构参数包括 β_1 和 β_2 。具体地，目标视频图像提取到的视频图像特征通过 $1 \times d \times d$ 的 2D 卷积进行空间特征提取，提取结果通过空间结构参数 α_1 进行贡献调整，具体通过特征提取结果

与结构参数进行相乘以进行融合，实现贡献调整，贡献调整后依次进行批量标准化处理和激活函数的非线性映射。映射后的结果分别通过两个 $t \times 1 \times 1$ 的 1D 卷积进行时间特征提取，提取的结果分别通过结构参数 β_1 和 β_2 进行加权融合，加权融合的结果通过时间结构参数 α_2 进行时间特征贡献调整，得到进行视频行为识别处理的行为识别特征。

具体地，在训练结构参数时，对基于微分方式更新的处理，预先定义一个多维的结构参数，如可以为多维的结构参数向量，具体为二维向量，在微分方式更新处理中具有梯度。其中，结构参数的维度分别代表空间卷积和时间卷积对应的结构参数。将结构参数作用于空间卷积和时间卷积来融合两者的特征，具体通过 α_1 作用于空间卷积进行贡献调整，通过 α_2 作用于时间卷积进行贡献调整。根据视频行为识别模型的预测结果和真实结果计算误差值，利用梯度下降算法对结构参数进行更新，在训练结束时得到训练完成的结构参数。

进一步地，在根据视频行为识别模型的预测结果和真实结果计算误差值，利用梯度下降算法对结构参数进行更新时，采用微分的方式进行优化。将网络结构搜索技术中操作空间记作 O ， o 则是具体的一个操作，节点指的是网络结构搜索方法中的基本操作单元的集合，设定 i 和 j 是两个顺序相邻的节点，它们之间的一组候选操作的权重记作 α_{ij} ， P 是对应的概率分布。节点 i 和 j 之间的具有最大概率的候选操作通过 $\max$ 函数得到，最终的网络结构通过不同节点间搜索得到的操作堆叠而成，如下式 (1) 所示，

$$\sum_{i=0, j=0}^N \max(P_{ij}(O_{ij})) \quad (1)$$

其中， N 为节点数量。

横向来看相当于学习选定的具体操作，将操作空间限制在级联的 2D 卷积和 1D 卷积之上直接通过梯度进行优化，以搜索得到对应的网络结构，如下式 (2)，

$$\nabla_w L_{train}(w, \alpha) \quad (2)$$

其中， ∇_w 为梯度优化处理， $L_{train}(w, \alpha)$ 为网络结构的目标函数， w 为网络结构的模型参数。

纵向来看, 相当于通过结构参数增强或者减弱 2D 空间卷积和 1D 时间卷积的特征在特征学习时的重要度。如图 6 所示, 本实施例的块定义在两个节点之间。例如, 对于 ResNet 结构这些节点代表前一个块的输出和后一个块的输入。顺序连接的 $1 \times d \times d$ 卷积和 $t \times 1 \times 1$ 卷积定义在块内部。结构参数用于这两个卷积之上来调整它们的强度。通过训练以从 $\alpha_{i1} \cdots \alpha_{in}$ 中寻找满足 2D 卷积的贡献调整要求的结构参数 α_1 , 从 $\alpha_{21} \cdots \alpha_{2j} \cdots \alpha_{2m}$ 中寻找满足 1D 卷积的贡献调整要求的结构参数 α_2 , 图 6 中确定 α_{in} 为结构参数 α_1 , α_{21} 为结构参数 α_2 。记 $o(\cdot)$ 为定义在搜索空间 O 中, 且作用于输入 x 之上的操作, 则节点 i 和节点 j 之间的权重向量是 $\alpha^{(i,j)}$, 可以得到下式 (3),

$$y^{(i,j)} = \sum_{o \in O} F_{i,j} \left(w_o, \alpha_o^{(i,j)} \right) o(x) \quad (3)$$

其中, F 是权重向量的线性映射, $y^{(i,j)}$ 是搜索空间中所有权重向量的线性映射之和, 具体可以将 F 设置为一个全连接层, 每一 cell 单元被定义为一个 (2+1)D 卷积块, 因此 $\alpha_o^{(i,j)}$ 是固定的。因此学习目标可以进一步简化为下式 (4),

$$y = g \left(w_\alpha, w_n, x \right) \quad (4)$$

其中, w_α 是网络的结构参数, w_n 是网络的模型参数, y 是 (2+1)D 卷积块的输出。得益于轻量的搜索空间, 具体实现时可以将结构参数和模型参数同时进行端到端的训练, 为每一个 (2+1)D 卷积块都学习一组结构参数, 从而得到的优化方式如下式 (5),

$$\nabla L_{val} \left(w_\alpha, w_n \right) \quad (5)$$

即对网络的结构参数 w_α 和模型参数 w_n 进行同步训练, 基于目标函数 L_{val} 进行梯度下降优化, 以得到满足需要的结构参数 w_α 和模型参数 w_n , 实现网络训练。

对于策略梯度方式的更新处理, 预先定义一个多维的结构参数, 如可以为多维的结构参数向量, 具体为二维向量, 在策略梯度方式的更新处理中截断梯度信息。其中, 结构参数的维度分别代表空间卷积和时间卷积对应的结构参数。预先定义一策略梯度 Agent 网络来根据

当前的结构参数和策略梯度 Agent 网络的网络状态生成下一个结构参数。将生成的结构参数作用于空间卷积和时间卷积来融合两者的特征。根据策略梯度 Agent 网络的当前网络状态的奖励值更新 Agent 的网络参数，进而由新的 Agent 来预测下一个结构参数，从而实现结构参数的更新。

具体地，策略梯度下降是一种强化学习方法，其中策略(policy)指的是在不同状态(state)下，所采取的动作(action)，目标是希望基于策略来做梯度下降，以此训练出策略梯度网络 Agent 能有较好的根据当前状态做出对应动作，能得到更高的奖励值(reward)。通过策略梯度方式来优化结构参数时，可以使用多层感知机(Multilayer Perceptron)作为策略梯度网络 Agent，当前策略梯度网络 Agent 的参数作为状态 state，网络输出的结构参数作为动作 action，使用当前主干网络，即使用视频行为识别模块的损失 loss 和奖励常量作为奖励值 reward 函数的组成部分。在前向的处理流程中，先输入初始的结构参数给 Agent 网络，紧接着该网络会根据当前 Agent 网络参数和输入的结构参数预测出下一个网络参数即 action。在反向传播过程中，则是最大化当前能够获得的奖励值 reward，通过奖励值对 Agent 网络的参数进行更新。设当前的状态是 s，a 代表当前的 action， θ 代表网络的参数，则交叉熵损失 CE 如下式 (6)，

$$CE(\hat{y}) = - \left[y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \right] \quad (6)$$

其中， $\hat{y}$ 为模型预测输出，y 为真实标签。为了确保结构参数搜索对网络整体学习的影响是正向的，可以基于平滑后的 CE 值进行 reward 函数的设计，可以使得搜索到的结构参数跟视频行为识别模型的主干网路的学习是互相辅助的。平滑后的 CE 如下式 (7)，

$$SCE = -(1 - \varepsilon)CE(i) + \varepsilon \sum \frac{CE(j)}{N} \quad (7)$$

其中， i ， j 和 N 分别为正确类别，其他类别和总类别数， ε 是一个非常小的常数。进一步地，如果后一个时间步 n 得到的 SCE_n 值大于前一个 m 得到的 SCE_m ，则给予正向的 reward

值 γ ，否则 reward 为 $-\gamma$ 。如下式 (8)，

$$f = -\gamma * \text{sgn}(SCE_m - SCE_n) \quad (8)$$

其中， f 为奖励值， γ 为设定的变量。

整体的目标函数如下式 (9)， $f(s, a)$ 为网络预测输出。

$$L = \sum \log_{\pi}(a|s, \theta) f(s, a) \quad (9)$$

具体地，针对时空信息重要度和缩小类内差异性的先验激励模块两部分的结构参数对应的多层感知机 MLP 分别是具有 6 个隐层神经元和 4 个隐层神经元的 3 层神经网络，同时在各层之间添加了 ReLU 激活函数，且最后一层为 softplus 激活函数。由于 policy gradient 机制需要完整的状态行为序列，则会使得中间状态缺少反馈进而导致整体训练效果不佳，因此对于状态序列长度，一种方法可以将其设置为 1 个时期 epoch，即每 2 个时期 epoch 计算一次最近时期 epoch 的 reward；另一种则可以将其看为一个迭代 (iteration) 内的优化，这样会更有利于优化。在优化时将网络的参数和 Agent 的参数进行剥离，分开优化。针对两种参数可以采用不同的优化器，其中 Agent 优化器采用 Adam 优化器，网络参数优化采用随机梯度下降 (Stochastic Gradient Descent, SGD) 进行优化处理，在优化时两者交替更新。

在将结构参数作用于空间卷积和时间卷积来融合两者的特征时，具体根据结构参数，使用 Auto(2+1)D 卷积结构，即 2D 卷积+1D 卷积的结构将视频图像特征中的时空信息进行融合。其中，Auto(2+1)D 是由顺序连接的 2D 卷积和 1D 卷积、各自对应的结构参数，以及激活函数组成。通过 2D 卷积和 1D 卷积来分别解耦特征中的时间和空间信息，进行独立建模，即通过 2D 卷积进行空间特征提取，通过 1D 卷积进行时间特征提取。在训练结构参数时，通过结构参数来自适应地对解耦后的信息进行融合，并通过激活函数增加模型的非线性表达能力。2D 卷积和 1D 卷积组成一个基本的卷积块，可以作为网络中的基础块结构，如可以作为 ResNet (Residual Neural Network, 残差网络) 中的 Block 结构。

在根据所提特征在时间维度的相似度和先验信息对应的结构参数，具体包括先验参数 β_1 和 β_2 ，使用节奏调节器调整行为的节奏的处理过程中，节奏调节器包含先验激励模块和高内聚时间表达模块。先验激励模块可以根据特征的时间维度上的相似度来为当前的结构参数设置界限值 **Margin**，以此促进结构参数的优化。高内聚的时间表达模块可以通过高效的注意力机制来增加时间维度信息的内聚性。具体地，将上一层输出的特征图输入 2D 卷积，进行空间特征的提取。将 2D 卷积输出的特征输入先验激励模块，计算其在时间维度上的相似度，并根据相似度值为结构参数设置合适的 **Margin**。另一方面，将 2D 卷积输出的特征输入高内聚时间模块和 1D 时间卷积模块并输出特征图，根据先验信息结构参数，自适应地调整高内聚时间模块和 1D 时间卷积模块输出的特征图的权重并进行融合，获得融合后的特征。

具体地，为了实现通过先验信息激励网络朝着高内聚时间特征的方向优化，将 $3 \times 1 \times 1$ 这一条时间卷积分支改为 $3 \times 1 \times 1$ 时间卷积和带有期望最大化注意力的 $3 \times 1 \times 1$ 时间卷积两个分支。先验激励模块主要是通过对先验参数 β_1 和 β_2 优化的激励作用于特征。如图 7 所示，目标视频图像提取到的视频图像特征通过 $1 \times 3 \times 3$ 的 2D 卷积进行空间特征提取，提取结果通过 α_1 进行贡献调整，贡献调整后依次进行批量标准化处理和激活函数的非线性映射。映射后的结果通过先验激励模块进行处理。在先验激励模块中，计算映射后的结果在时间维度的相似度，基于相似度对初始的先验参数 β_1 和 β_2 进行修正，并通过修改后的先验参数 β_1 和 β_2 ，对通过两个 $t \times 1 \times 1$ 的 1D 卷积进行时间特征提取得到的结果进行加权融合，加权融合的结果通过结构参数 α_2 进行时间特征贡献调整，得到进行视频行为识别处理的行为识别特征。

图 7 中，箭头代表特征图的流向，它们通过把上一个模块输出的特征图输入下一个模块的方式连接，紧接着将先验相似度激励模块后得到的特征图，并行的输入下一个卷积块，最后的输出是将两个分支的特征图进行拼接并降低维度。为了通过先验信息激励网络朝着高内聚时间特征的或者高静态特性的方向优化，首先根据 $1 \times 3 \times 3$ 卷积得到的特征图在时间维度上计算余弦相似度，以此来衡量该样本在时间维度的变化程度，并基于该变化程度阈值将当前

的先验参数划分为正负参数。在具体实现中，对于动作节奏较慢的视频，各帧目标视频图像之间的冗余信息多，则需要增强内聚特征，可以增大内聚特征的权重，以突出焦点特征进行行为识别，从而提高视频行为识别的准确度。具体来说，经过激励修正后的先验参数与原始输入的先验参数以残差连接的方式合并作为最终的先验参数。由于在网络达到一定优化的情况下，张量的元素值往往没有较大的方差，统一偏小，在具体实现时可以通过设置界限值 *margin*，动态地调整阈值来设置当前的相似度先验信息，可以获得如下式 (10)，

$$Sim = \max\left(0, Sim - \left(Thres + abs(\beta_1 - \beta_2)\right)\right) \quad (10)$$

其中，*Sim* 代表相似度值，*Thres* 是阈值， β_1 和 β_2 为先验参数。

进一步地，高内聚时间模块基于 EM (Expectation-Maximum，称期望最大化) 算法优化的注意力机制获得高内聚的时间表达。对于每个样本，都经过固定次数的迭代优化来重构特征。如图 8 所示，这个过程可以分为 E 步和 M 步，特征图经过下采样处理后，通过 E 步和 M 步分别进行处理后融合得到高内聚特征。首先，假设有基向量 $B \times C \times K$ ，其中 *B* 为 batch 大小，即为批次处理的数据大小，*C* 为原始输入的视频图像特征对应的通道数，*K* 为基向量维度。在 E 步中，通过使用基向量和 $B \times (H \times W) \times C$ 的进行空间特征提取后的空间特征向量做矩阵乘法，再接 softmax 来重构原始特征，得到尺寸为 $B \times (H \times W) \times K$ 的特征图。在 M 步中则是将尺寸为 $B \times (H \times W) \times K$ 的重构特征图和 $B \times (H \times W) \times C$ 的原始特征图做乘法来得到新的基向量 $B \times C \times K$ 。进一步地，为了保证基向量更新的稳定，对其进行 L2 正则化，同时在训练时增加基向量的滑动平均更新，具体如下式 (11)，

$$mu = mu * momentum + mu_mean * (1 - momentum) \quad (11)$$

其中，*mu* 为基向量，*mu_mean* 为其均值，*momentum* 为动量。

最后将 E 步得到的基向量和 M 步得到的注意力图做矩阵乘法，得到最终重构的带有全局信息的特征图。

本实施例中提供的视频行为识别方法应用于视频识别领域，而在视频识别领域中三维卷

积目前被广泛的使用,但是由于其参数量高的限制难以拓展。一些改进方法在计算成本低、内存需求小、性能高的基础上,将三维卷积分解为二维空间卷积和一维时间卷积。随后很多工作着力于通过设计不同的网络结构来获得更具有表达性的特征。但业内并未关注到视频中的空间和时间线索在不同的动作类别上有着不同的影响。而本实施例中的视频行为识别方法,涉及的自适应时空纠缠网络基于重要性分析自动地融合分解后的时空信息,以获得更强大的时空表示。该视频行为识别方法中,Auto (2+1) D 卷积通过网络结构搜索技术自适应重组解耦时空卷积滤波器,以建模时空的不一致贡献信息,挖掘出了时空信息之间的深层相关性,并学习时空交互信息,通过整合不同权重的时空信息,增强了当前模型对时间和空间信息的建模能力。而节律调节器利用 EM 算法的有效注意机制来提取时间维度的高内聚特征,可以根据动作节奏的先验信息和时间卷积的结构参数,来调整具有不同节奏的动作的时间信息,以此获得时间信息的高内聚性的表达式来处理不同动作类中的不同持续时间问题,可以提高视频行为识别的准确率。

应该理解的是,虽然图 2-图 3 的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,图 2-图 3 中的至少一部分步骤可以包括多个步骤或者多个阶段,这些步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

在一个实施例中,如图 9 所示,提供了一种视频行为识别装置 900,该装置可以采用软件模块或硬件模块,或者是二者的结合成为计算机设备的一部分,该装置具体包括:视频图像特征提取模块 902、空间特征贡献调整模块 904、特征融合模块 906、时间特征贡献调整模块 908 和视频行为识别模块 910,其中:

视频图像特征提取模块 902,用于从至少两帧目标视频图像提取视频图像特征;

空间特征贡献调整模块 904,用于将视频图像特征的空间特征进行贡献调整,得到中间图像特征;

特征融合模块 906,用于基于先验信息对中间图像特征的时间特征和时间特征对应的内聚特征进行融合,得到融合特征;先验信息是根据中间图像特征在时间维度的变化信息得到的;内聚特征是对时间特征进行关注处理得到的;

时间特征贡献调整模块 908,用于对融合特征进行时间特征贡献调整,得到行为识别特征;

视频行为识别模块 910,用于基于行为识别特征进行视频行为识别。

在一个实施例中,空间特征贡献调整模块 904,还用于将视频图像特征进行空间特征提取,得到视频图像特征的空间特征;通过结构参数中的空间结构参数对空间特征进行贡献调整,得到中间图像特征;结构参数是通过携带行为标签的视频图像样本训练得到的;时间特征贡献调整模块 908,还用于通过结构参数中的时间结构参数对融合特征进行贡献调整,得到行为识别特征。

在一个实施例中,还包括待训练参数确定模块、中间样本特征获得模块、融合样本特征获得模块、行为识别样本特征获得模块和迭代模块;其中:待训练参数确定模块,用于确定待训练结构参数;中间样本特征获得模块,用于通过待训练结构参数中的空间结构参数,对视频图像样本特征的空间样本特征进行贡献调整,得到中间样本特征;视频图像样本特征是从视频图像样本提取得到的;融合样本特征获得模块,用于基于先验样本信息对中间样本特征的时间样本特征和时间样本特征对应的内聚样本特征进行融合,得到融合样本特征;内聚样本特征是对时间样本特征进行关注处理得到的;先验样本信息是根据中间样本特征在时间维度的变化信息得到的;行为识别样本特征获得模块,用于通过待训练结构参数中的时间结构参数对融合样本特征进行贡献调整,得到行为识别样本特征;迭代模块,用于基于行为识别样本特征进行视频行为识别,并根据行为识别结果和视频图像样本对应的行为标签,对待

训练结构参数进行更新后继续训练直至训练结束，获得结构参数。

在一个实施例中，视频行为识别装置通过视频行为识别模型实现，待训练结构参数是视频行为识别模型在训练中的参数；迭代模块还包括识别结果获得模块、差异确定模块、结构参数更新模块和结构参数获得模块；其中：识别结果获得模块，用于获得视频行为识别模型输出的行为识别结果；差异确定模块，用于确定行为识别结果与视频图像样本对应的行为标签之间的差异；结构参数更新模块，用于根据差异对视频行为识别模型中的模型参数和待训练结构参数进行更新；结构参数获得模块，用于基于更新后的视频行为识别模型继续训练直至训练结束，并根据训练完成的视频行为识别模型得到结构参数。

在一个实施例中，迭代模块还包括识别损失确定模块、奖励值获得模块和奖励值处理模块；其中：识别损失确定模块，用于确定行为识别结果和视频图像样本对应的行为标签之间的行为识别损失；奖励值获得模块，用于根据行为识别损失和前一行为识别损失得到奖励值；奖励值处理模块，用于根据奖励值对待训练结构参数进行更新，通过更新后的待训练结构参数继续训练直至目标函数满足结束条件时，获得结构参数；目标函数基于训练过程中的各奖励值得到。

在一个实施例中，奖励值获得模块，还用于根据奖励值对策略梯度网络模型的模型参数进行更新；由更新后的策略梯度网络模型对待训练结构参数进行更新。

在一个实施例中，奖励值获得模块，还用于通过更新后的策略梯度网络模型，基于更新后的模型参数和待训练结构参数进行结构参数预测，获得预测的结构参数；及根据预测的结构参数，得到对待训练结构参数进行更新后的结构参数。

在一个实施例中，还包括相似度确定模块和先验信息修正模块；其中：相似度确定模块，用于确定中间图像特征在时间维度的相似度；先验信息修正模块，用于基于相似度对初始先验信息进行修正，得到先验信息。

在一个实施例中，初始先验信息包括第一初始先验参数和第二初始先验参数；先验信息修正模块包括相似度调整模块、先验参数修正模块和先验信息获得模块；其中：相似度调整模块，用于根据第一初始先验参数、第二初始先验参数及预设阈值，对相似度进行动态调整；先验参数修正模块，用于通过动态调整后的相似度分别对第一初始先验参数和第二初始先验参数进行修正，得到第一先验参数和第二先验参数；先验信息获得模块，用于根据第一先验参数和第二先验参数得到先验信息。

在一个实施例中，还包括基向量确定模块、特征重构模块、基向量更新模块和内聚特征获得模块；其中：基向量确定模块，用于确定当前基向量；特征重构模块，用于通过当前基向量对中间图像特征的时间特征进行特征重构，得到重构特征；基向量更新模块，用于根据重构特征和时间特征生成下一关注处理的基向量；内聚特征获得模块，用于根据下一关注处理的基向量、基向量和时间特征，得到时间特征对应的内聚特征。

在一个实施例中，基向量更新模块还包括注意力特征模块、正则化处理模块和滑动平均更新模块；其中：注意力特征模块，用于融合重构特征和时间特征，生成注意力特征；正则化处理模块，用于对注意力特征进行正则化处理，得到正则化特征；滑动平均更新模块，用于对正则化特征进行滑动平均更新，生成下一关注处理的基向量。

在一个实施例中，当前基向量包括批次处理的数据大小、中间图像特征的通道数以及基向量的维度；特征重构模块，还用于将当前基向量与中间图像特征的时间特征，依次进行矩阵相乘及归一化映射处理，得到重构特征；基向量更新模块，还用于将重构特征和时间特征进行矩阵相乘，得到下一关注处理的基向量；内聚特征获得模块，还用于将下一关注处理的基向量、基向量和时间特征进行融合，得到时间特征对应的内聚特征。

在一个实施例中，特征融合模块 906，还用于确定先验信息；对中间图像特征进行时间特征提取，得到中间图像特征的时间特征；通过先验信息，对时间特征和时间特征对应的内聚特征进行加权融合，得到融合特征。

在一个实施例中，先验信息包括第一先验参数和第二先验参数；特征融合模块 906，还

用于通过第一先验参数对时间特征进行加权处理，获得加权处理后的时间特征；通过第二先验参数对时间特征对应的内聚特征进行加权处理，得到加权处理后的内聚特征；及将加权处理后的时间特征和加权处理后的内聚特征进行融合，得到融合特征。

在一个实施例中，还包括标准化处理模块和非线性映射模块；其中：标准化处理模块，用于对中间图像特征进行标准化处理，得到标准化特征；非线性映射模块，用于根据标准化特征进行非线性映射，获得映射后的中间图像特征；特征融合模块 906，还用于基于先验信息对映射后的中间图像特征的时间特征和时间特征对应的内聚特征进行融合，得到融合特征；先验信息是根据映射后的中间图像特征在时间维度的变化信息得到的。

在一个实施例中，标准化处理模块，还用于通过批量标准化层结构，对中间图像特征进行标准化处理，得到标准化特征；非线性映射模块，还用于通过激活函数对标准化特征进行非线性映射，获得映射后的中间图像特征。

关于视频行为识别装置的具体限定可以参见上文中对于视频行为识别方法的限定。上述视频行为识别装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器或终端，其内部结构图可以如图 10 所示。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储模型数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种视频行为识别方法。

本领域技术人员可以理解，图 10 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

在一个实施例中，还提供了一种计算机设备，包括存储器和处理器，存储器中存储有计算机可读指令，该处理器执行计算机可读指令时实现上述各方法实施例中的步骤。

在一个实施例中，提供了一种计算机可读存储介质，存储有计算机可读指令，该计算机可读指令被处理器执行时实现上述各方法实施例中的步骤。

在一个实施例中，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机可读指令，该计算机可读指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机可读指令，处理器执行该计算机可读指令，使得该计算机设备执行上述各方法实施例中的步骤。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性计算机可读存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器（Read-Only Memory, ROM）、磁带、软盘、闪存或光存储器等。易失性存储器可包括随机存取存储器（Random Access Memory, RAM）或外部高速缓冲存储器。作为说明而非局限，RAM 可以是多种形式，比如静态随机存取存储器（Static Random Access Memory, SRAM）或动态随机存取存储器（Dynamic Random Access Memory, DRAM）等。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因

此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

- 1、一种视频行为识别方法，由计算机设备执行，其特征在于，所述方法包括：
从至少两帧目标视频图像提取视频图像特征；
将所述视频图像特征的空间特征进行贡献调整，得到中间图像特征；
基于先验信息对所述中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征；所述先验信息是根据所述中间图像特征在时间维度的变化信息得到的；所述内聚特征是对所述时间特征进行关注处理得到的；
对所述融合特征进行时间特征贡献调整，得到行为识别特征；及
基于所述行为识别特征进行视频行为识别。
- 2、根据权利要求1所述的方法，其特征在于，所述将所述视频图像特征的空间特征进行贡献调整，得到中间图像特征，包括：
将所述视频图像特征进行空间特征提取，得到所述视频图像特征的空间特征；及
通过结构参数中的空间结构参数对所述空间特征进行贡献调整，得到中间图像特征；所述结构参数是通过携带行为标签的视频图像样本训练得到的；
所述对所述融合特征进行时间特征贡献调整，得到行为识别特征，包括：
通过所述结构参数中的时间结构参数对所述融合特征进行贡献调整，得到行为识别特征。
- 3、根据权利要求2所述的方法，其特征在于，所述方法还包括：
确定待训练结构参数；
通过所述待训练结构参数中的空间结构参数，对视频图像样本特征的空间样本特征进行贡献调整，得到中间样本特征；所述视频图像样本特征是从所述视频图像样本中提取得到的；
基于先验样本信息对所述中间样本特征的时间样本特征和所述时间样本特征对应的内聚样本特征进行融合，得到融合样本特征；所述内聚样本特征是对所述时间样本特征进行关注处理得到的；所述先验样本信息是根据所述中间样本特征在时间维度的变化信息得到的；
通过所述待训练结构参数中的时间结构参数对所述融合样本特征进行贡献调整，得到行为识别样本特征；及
基于所述行为识别样本特征进行视频行为识别，并根据行为识别结果和所述视频图像样本对应的行为标签，对所述待训练结构参数进行更新后继续训练直至训练结束，获得所述结构参数。
- 4、根据权利要求3所述的方法，其特征在于，所述方法通过视频行为识别模型实现，所述待训练结构参数是所述视频行为识别模型在训练中的参数；所述根据行为识别结果和所述视频图像样本对应的行为标签，对所述待训练结构参数进行更新后继续训练直至训练结束，获得所述结构参数，包括：
获得所述视频行为识别模型输出的行为识别结果；
确定所述行为识别结果与所述视频图像样本对应的行为标签之间的差异；
根据所述差异对所述视频行为识别模型中的模型参数和所述待训练结构参数进行更新；
及
基于更新后的视频行为识别模型继续训练直至训练结束，并根据训练完成的视频行为识别模型得到所述结构参数。
- 5、根据权利要求3所述的方法，其特征在于，所述根据行为识别结果和所述视频图像样本对应的行为标签，对所述待训练结构参数进行更新后继续训练直至训练结束，获得所述结构参数，包括：
确定行为识别结果和所述视频图像样本对应的行为标签之间的行为识别损失；
根据所述行为识别损失和前一行为识别损失得到奖励值；及
根据所述奖励值对所述待训练结构参数进行更新，通过更新后的待训练结构参数继续训练直至目标函数满足结束条件时，获得所述结构参数；所述目标函数基于训练过程中的各奖励值得到。

6、根据权利要求5所述的方法，其特征在于，所述根据所述奖励值对所述待训练结构参数进行更新，包括：

根据所述奖励值对策略梯度网络模型的模型参数进行更新；及
由更新后的策略梯度网络模型对所述待训练结构参数进行更新。

7、根据权利要求6所述的方法，其特征在于，所述由更新后的策略梯度网络模型对所述待训练结构参数进行更新，包括：

通过更新后的策略梯度网络模型，基于更新后的模型参数和待训练结构参数进行结构参数预测，获得预测的结构参数；及

根据所述预测的结构参数，得到对所述待训练结构参数进行更新后的结构参数。

8、根据权利要求1所述的方法，其特征在于，所述方法还包括：

确定所述中间图像特征在时间维度的相似度；及

基于所述相似度对初始先验信息进行修正，得到先验信息。

9、根据权利要求8所述的方法，其特征在于，所述初始先验信息包括第一初始先验参数和第二初始先验参数；所述基于所述相似度对初始先验信息进行修正，得到先验信息，包括：

根据所述第一初始先验参数、所述第二初始先验参数及预设阈值，对所述相似度进行动态调整；

通过动态调整后的相似度分别对所述第一初始先验参数和所述第二初始先验参数进行修正，得到第一先验参数和第二先验参数；及

根据所述第一先验参数和所述第二先验参数得到先验信息。

10、根据权利要求1所述的方法，其特征在于，所述方法还包括：

确定当前基向量；

通过所述当前基向量对所述中间图像特征的时间特征进行特征重构，得到重构特征；

根据所述重构特征和所述时间特征生成下一关注处理的基向量；及

根据所述下一关注处理的基向量、所述基向量和所述时间特征，得到所述时间特征对应的内聚特征。

11、根据权利要求10所述的方法，其特征在于，所述根据所述重构特征和所述时间特征生成下一关注处理的基向量，包括：

融合所述重构特征和所述时间特征，生成注意力特征；

对所述注意力特征进行正则化处理，得到正则化特征；及

对所述正则化特征进行滑动平均更新，生成下一关注处理的基向量。

12、根据权利要求10所述的方法，其特征在于，所述当前基向量包括批次处理的数据大小、中间图像特征的通道数以及基向量的维度；所述通过所述当前基向量对所述中间图像特征的时间特征进行特征重构，得到重构特征，包括：

将所述当前基向量与所述中间图像特征的时间特征，依次进行矩阵相乘及归一化映射处理，得到重构特征；

所述根据所述重构特征和所述时间特征生成下一关注处理的基向量，包括：

将所述重构特征和所述时间特征进行矩阵相乘，得到下一关注处理的基向量；

所述根据所述下一关注处理的基向量、所述基向量和所述时间特征，得到所述时间特征对应的内聚特征，包括：

将所述下一关注处理的基向量、所述基向量和所述时间特征进行融合，得到所述时间特征对应的内聚特征。

13、根据权利要求1至12任意一项所述的方法，其特征在于，所述基于先验信息对所述中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征，包括：

确定先验信息；

对所述中间图像特征进行时间特征提取，得到所述中间图像特征的时间特征；及

通过所述先验信息，对所述时间特征和所述时间特征对应的内聚特征进行加权融合，得

到融合特征。

14、根据权利要求 13 所述的方法，其特征在于，所述先验信息包括第一先验参数和第二先验参数；所述通过所述先验信息，对所述时间特征和所述时间特征对应的内聚特征进行加权融合，得到融合特征，包括：

通过所述第一先验参数对所述时间特征进行加权处理，获得加权处理后的时间特征；

通过所述第二先验参数对所述时间特征对应的内聚特征进行加权处理，得到加权处理后的内聚特征；及

将所述加权处理后的时间特征和所述加权处理后的内聚特征进行融合，得到融合特征。

15、根据权利要求 1 所述的方法，其特征在于，在所述基于先验信息对所述中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征之前，还包括：

对所述中间图像特征进行标准化处理，得到标准化特征；及

根据所述标准化特征进行非线性映射，获得映射后的中间图像特征；

所述基于先验信息对所述中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征，包括：

基于先验信息对所述映射后的中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征；所述先验信息是根据所述映射后的中间图像特征在时间维度的变化信息得到的。

16、根据权利要求 15 所述的方法，其特征在于，所述对所述中间图像特征进行标准化处理，得到标准化特征，包括：

通过批量标准化层结构，对所述中间图像特征进行标准化处理，得到标准化特征；

所述根据所述标准化特征进行非线性映射，获得映射后的中间图像特征，包括：

通过激活函数对所述标准化特征进行非线性映射，获得映射后的中间图像特征。

17、一种视频行为识别装置，其特征在于，所述装置包括：

视频图像特征提取模块，用于从至少两帧目标视频图像提取视频图像特征；

空间特征贡献调整模块，用于将所述视频图像特征的空间特征进行贡献调整，得到中间图像特征；

特征融合模块，用于基于先验信息对所述中间图像特征的时间特征和所述时间特征对应的内聚特征进行融合，得到融合特征；所述先验信息是根据所述中间图像特征在时间维度的变化信息得到的；所述内聚特征是对所述时间特征进行关注处理得到的；

时间特征贡献调整模块，用于对所述融合特征进行时间特征贡献调整，得到行为识别特征；及

视频行为识别模块，用于基于所述行为识别特征进行视频行为识别。

18、一种计算机设备，包括存储器和处理器，所述存储器存储有计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现权利要求 1 至 16 中任一项所述的方法的步骤。

19、一种计算机可读存储介质，存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现权利要求 1 至 16 中任一项所述的方法的步骤。

20、一种计算机程序产品，包括计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现权利要求 1 至 16 任一项所述的方法的步骤。