

动作识别方法、装置、设备及存储介质

本申请要求在 2020 年 5 月 20 日提交中国专利局、申请号为 202010431706.9 的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及计算机视觉应用技术领域，例如涉及动作识别方法、装置、设备及存储介质。

背景技术

基于视频的动作识别，一直是计算机视觉研究的重要领域。视频动作识别的实现主要包括特征抽取与表示，以及特征分类两大部分。经典的如密度轨迹追踪等方法，一般为手动设计特征的方法，而近些年来，人们发现深度学习具备强大的特征表示能力，神经网络便逐渐成为动作识别领域的主流方法，相对于手动设计特征的方法，大大提升了动作识别的性能。

目前，神经网络动作识别方案大部分基于从视频中获取的序列图片构建出时序关系，从而对动作进行判断。例如：基于循环神经网络构建图片之间的时序、基于 3D 卷积提取多个图片时间的时序信息以及基于图片的深度学习技术叠加动作变化的光流信息等。上述方案存在计算量与识别精度不能兼顾的情况，需要改进。

发明内容

以下是对本文详细描述的主题的概述。本概述并非是为了限制权利要求的保护范围。

本申请实施例提供了动作识别方法、装置、设备及存储介质，可以优化相关技术中的针对视频的动作识别方案。

第一方面，本申请实施例提供了一种动作识别方法，该方法包括：

对原始压缩视频数据进行分组处理，得到分组视频数据；

将所述分组视频数据输入至第一预设模型中，并根据所述第一预设模型的输出结果确定包含动作的目标分组视频数据；

对所述目标分组视频数据进行解码，得到待识别分组视频数据；

将所述待识别分组视频数据输入至第二预设模型中，并根据所述第二预设模型的输出结果确定所述待识别分组视频数据中包含的动作类型。

第二方面，本申请实施例提供了一种动作识别装置，该装置包括：

视频分组模块，设置为对原始压缩视频数据进行分组处理，得到分组视频数据；

目标分组视频确定模块，设置为将所述分组视频数据输入至第一预设模型中，并根据所述第一预设模型的输出结果确定包含动作的目标分组视频数据；

视频解码模块，设置为对所述目标分组视频数据进行解码，得到待识别分组视频数据；

动作类型识别模块，设置为将所述待识别分组视频数据输入至第二预设模型中，并根据所述第二预设模型的输出结果确定所述待识别分组视频数据中包含的动作类型。

第三方面，本申请实施例提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如本申请实施例提供的动作识别方法。

第四方面，本申请实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本申请实施例提供的动作识别方法。

附图说明

图1为本申请实施例提供的一种动作识别方法的流程示意图；

图2为本申请实施例提供的一种压缩视频中帧排列示意图；

图3为本申请实施例提供的一种特征变换操作示意图；

图4为本申请实施例提供的另一种动作识别方法的流程示意图；

图5为本申请实施例提供的一种基于压缩视频流的动作识别过程示意图；

图6为本申请实施例提供的一种第一2D残差网络结构示意图；

图7为本申请实施例提供的一种动作标签示意图；

图8为本申请实施例提供的又一种动作识别方法的流程示意图；

图9为本申请实施例提供的基于短视频的动作识别应用场景示意图；

图10为本申请实施例提供的一种基于序列图片的动作识别过程示意图；

图11为本申请实施例提供的一种第二2D残差网络结构示意图；

图 12 为本申请实施例提供的一种 3D 残差网络结构示意图；

图 13 为本申请实施例提供的两级多感受野池化操作的运算过程示意图；

图 14 为本申请实施例提供的一种动作识别装置的结构框图；

图 15 为本申请实施例提供的一种计算机设备的结构框图。

具体实施方式

下面结合附图和实施例对本申请作详细说明。可以理解的是，此处所描述的示例实施例仅仅用于解释本申请，而非对本申请的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本申请相关的部分而非全部结构。此外，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

本申请实施例中的动作识别方案可应用于各种针对视频的动作识别场景，如短视频审核场景、视频监控场景、实时通话识别场景以及机器人视觉识别场景等等。其中，视频可以是视频文件，也可以是视频流。

目前，神经网络动作识别方案大部分基于从视频中获取的序列图片构建出时序关系，从而对动作进行判断。例如：基于循环神经网络（Recurrent Neural Network, RNN）或长短期记忆网络（LSTM, Long Short-Term Memory）等构建图片之间的时序、基于 3D 卷积提取多个图片时间的时序信息以及基于图片的深度学习技术叠加动作变化的光流信息。

以短视频审核场景为例，基于序列图片进行动作识别的方法至少存在以下两个不足：第一，这些技术方案对一个短视频做出判断都需要大量的计算资源，重度依赖机器中央处理器（Central Processing Unit, CPU）以及图像处理器（Graphics Processing Unit, GPU）的计算力，而且从压缩的短视频中提取出图片需要经过解码，如 decode 解码（将压缩视频解码成图片的技术），而对短时视频进行全时长段解码本身又需要大量的 CPU 和 GPU 算力，因此基于图片的动作识别方案对机器算力需求高，且随着从短视频中取帧的间隔的减小，计算资源的需求与短视频时长成线性增长关系；第二，基于循环神经网络和光流的技术方案，机器审核精度较低，为保证对目标动作的召回率，势必增加机审推送比，导致人审阶段对人力需求变大，从而增大了审核成本。由此可见，上述方案存在计算量与识别精度不能兼顾的情况，需要改进，且除短视频审核场景外的其他类似应用场景同理也存在上述情况。

图 1 为本申请实施例提供的一种动作识别方法的流程示意图，该方法可以

由动作识别装置执行，其中该装置可由软件和/或硬件实现，一般可集成在计算机设备中。其中，计算机设备例如可以是服务器，也可以是手机等设备，也可由两种或两种以上的设备分别执行部分步骤。如图 1 所示，该方法包括：

步骤 101、对原始压缩视频数据进行分组处理，得到分组视频数据。

示例性的，视频数据中包含了大量的图像以及声音信息，在传输或存储时，通常需要对视频数据进行压缩编码，得到压缩视频数据，本申请实施例这里称为原始压缩视频数据。编码标准以及编码参数等不做限定，例如可以是 H264、H265 以及 MPEG-4 等等。

示例性的，可以基于预设分组规则对原始压缩视频数据进行区间划分，得到区间压缩视频（segment，以下可简称片段），将区间压缩视频作为分组视频数据，或从区间压缩视频中选取部分数据作为分组视频数据。其中，预设分组规则中例如可包含区间划分时间间隔，即被区间划分后的每个区间压缩视频对应的时长，区间划分时间间隔可以是恒定的，也可以是变化的，不做限定。以短视频审核场景为例，区间划分时间间隔例如可以是 5 秒钟。另外，预设分组规则中例如也可以包含区间数量，数值不做限定。

需要说明的是，虽然本文将区间压缩视频简称为片段，但仅为了描述方便，上述分组处理及区间划分可以不涉及对原始压缩视频进行切分或切割的操作，避免引入额外的计算力和存储，保证工程开发效率。分组处理及区间划分可以利用时间戳来完成，也即可以由起止时间戳来限定分组视频数据或区间压缩视频的区间范围。本步骤可理解为从原始压缩视频数据中提取到不同时间段的数据分别输入至第一预设模型。以区间划分时间间隔为 5 秒为例，原始压缩视频中 0~5s 对应的数据为一段分组视频数据，5~10s 对应的数据为一段分组视频数据，2 段数据是分别进入第一预设模型的。

步骤 102、将分组视频数据输入至第一预设模型中，并根据所述第一预设模型的输出结果确定包含动作的目标分组视频数据。

示例性的，对原始压缩视频数据进行分组处理时，可以每得到一段分组视频数据就输入至第一预设模型中，提高计算效率；也可以全部分组完成后，再依次或并行将各段分组视频数据输入至第一预设模型中，确保在分组处理准确完成的情况下再进行粗粒度的动作识别。

示例性的，第一预设模型可以是预先训练的神经网络模型，在需要使用时直接进行加载，该模型主要设置为识别分组视频数据中是否包含动作，并不关

心是哪个动作，可以设置动作标签为二分类，如“是”和“否”，可在第一预设模型的训练样本中标记二分类标签。这样，根据第一预设模型的输出结果就可以筛选出哪些片段中包含动作，将对应的分组视频数据确定为目标分组视频数据。由于不需要识别动作类型，也即仅进行粗粒度的识别，因此第一预设模型的计算量较小，且在未解码的情况下进行识别，可以节省大量的解码算力，在排除大量不包含动作的视频片段的同时保证包含动作的视频片段被保留，用于后续识别。

其中，第一预设模型的网络结构以及相关参数等本申请实施例不做限定，可根据实际需求设置，例如可以为轻量级的模型。

步骤 103、对所述目标分组视频数据进行解码，得到待识别分组视频数据。

示例性的，可参考原始压缩视频的编码标准等因素选择适当的解码方式，不做限定。在对目标分组视频数据进行解码后，可以将得到的视频图像作为待识别分组视频数据（解码后的视频图像一般按照时间顺序排列，即序列图像，此时待识别分组视频数据中可包含视频图像的时序信息），也可以在此基础上提取其他信息一并作为待识别分组视频数据，这里的其他信息例如可以是频域信息等。

步骤 104、将所述待识别分组视频数据输入至第二预设模型中，并根据所述第二预设模型的输出结果确定所述待识别分组视频数据中包含的动作类型。

示例性的，第二预设模型可以是预先训练的神经网络模型，在需要使用时直接进行加载，该模型主要设置为识别待识别分组视频数据中包含的动作的类型，也即进行细粒度的识别，可在第二预设模型的训练样本中标记多分类标签，这样，根据第二预设模型的输出结果就可以确定出最终的动作识别结果。经过前述步骤的初筛，能够进入到第二预设模型的数据相比原始压缩视频数据已经大大减少，且数据纯度也远高于原始压缩视频数据，需要识别的视频片段数量的量级不大，所以可以采用解码之后的序列图像进行识别。在一些实施例中，可采用神经网络参数较多的 3D 卷积网络结构进行时序特征的提取，标签则是采用粒度更细的多分类，标签数量不做限定，例如可以是 50 个，可根据的识别精度进行调整。

本申请实施例中提供的动作识别方法，对原始压缩视频数据进行分组处理，得到分组视频数据，将分组视频数据输入至第一预设模型中，并根据第一预设模型的输出结果确定包含动作的目标分组视频数据，对目标分组视频数据进行

解码，得到待识别分组视频数据，将待识别分组视频数据输入至第二预设模型中，并根据第二预设模型的输出结果确定待识别分组视频数据中包含的动作类型。通过采用上述技术方案，在对压缩视频进行解压前，先利用第一预设模型粗略筛选出包含动作的视频片段，再利用第二预设模型精确识别包含的动作的类型，可以在保证识别精度的前提下有效减少计算量，提高动作识别效率。

需要说明的是，对于一些如视频审核的应用场景，一般可包括用于视频上传的客户端和视频审核的服务器，由于上述步骤 101 和步骤 102 的相关操作计算量较小，可以在客户端完成，也即客户端可以针对即将上传的压缩视频进行初步筛选，将包含动作的目标分组视频数据上传至服务器进行识别和审核。另外，对于一些配置较高的设备来说，也可在客户端本地完成整个识别流程，即步骤 101 至步骤 104 的相关操作，实现根据最终识别出来的动作类型来确定是否允许视频的上传等控制。

在一些实施例中，所述对原始压缩视频数据进行分组处理，得到分组视频数据，包括：基于预设分组规则对原始压缩视频数据进行区间划分，得到区间压缩视频；基于预设提取策略提取所述区间压缩视频中的 I 帧数据和 P 帧数据，得到分组视频数据，其中，所述 P 帧数据包含 P 帧对应的运动矢量信息和/或色素变化残差信息。本申请实施例提供的技术方案，可以快速提取压缩视频数据中可以用于动作识别的特征数据，提高识别效率。

其中，I 帧又称关键帧，指压缩视频中包含的图像；P 帧又称前向预测编码帧，一般包括压缩视频中参照 I 帧的变化信息，可包括运动矢量 (motion vector, MV) 信息和 RGB 色素变化的残差 (RGB Residual Frame, RGBR) 信息。一个 I 帧后面通常存在多个 P 帧，不同编码方式中，I 帧和 P 帧的分布或包含的内容可能存在差异。示例性的，预设分组规则可如前文所述，例如可包括区间划分时间间隔或分段数量等。预设提取策略例如可包括 I 帧的提取策略和 P 帧的提取策略。I 帧的提取策略可包括获取 I 帧的时间间隔，或单位时间内获取的 I 帧数量，例如，每秒钟获取 1 个 I 帧。P 帧的提取策略可包括获取一个 I 帧后面的 P 帧的数量以及每两个被获取的 P 帧之间的时间间隔，例如数量为 2。那么以一个区间压缩视频为 5 秒为例，可以获取 5 个 I 帧和 10 个 P 帧对应的数据，若 P 帧数据同时包括 MV 和 RGBR，则可获取 10 个 MV 和 10 个 RGBR。

在一些实施例中，所述第一预设模型中包含第一 2D 残差网络、第一拼接层和第一全连接层；所述分组视频数据被输入至所述第一预设模型中后，经由所

述第一 2D 残差网络得到对应的维度相同的特征图；所述特征图经由所述第一拼接层得到按照帧的先后顺序进行拼接操作后的拼接特征图；所述拼接特征图经由所述第一全连接层得到是否包含动作的分类结果。本申请实施例提供的技术方案，采用比较精简的网络结果来获取较高的识别效率，并保证包含动作的视频片段有较高的召回率。

示例性的，第一 2D 残差网络可以采用轻量级的 ResNet18 模型。由于 I 帧、MV 和 RGBR 在数据特征上并不能直接拼接（concat）在一起，因此可以使用 3 个 ResNet18 对 I 帧、MV 和 RGBR 分别进行单独处理，得出维度相同的特征图，可分别记为 I 帧特征图、MV 特征图和 RGBR 特征图，MV 特征图和 RGBR 特征图统称 P 帧特征图。其中，维度相同可以指 $C \times H \times W$ 一致，其中，C 表示通道（channel）、H 表示高度（height）以及 W 表示宽度（width），* 又可表示为 $\times$ 。经过处理后，维度相同的特征图便可经过第一拼接层实现拼接。

在一些实施例中，基于预设提取策略提取所述区间压缩视频中的 I 帧数据和 P 帧数据，得到分组视频数据，包括：基于预设提取策略提取所述区间压缩视频中的 I 帧数据和 P 帧数据；对所述 P 帧数据进行累加变换，以使得变换后的 P 帧数据依赖于前向相邻的 I 帧；根据所述 I 帧数据和变换后的 P 帧数据确定分组视频数据。相应的，所述第一预设模型中还包括位于所述拼接层之前的相加层，所述特征图中对应 P 帧数据的特征图记为 P 帧特征图，所述特征图中对应 I 帧数据的特征图记为 I 帧特征图；所述 P 帧特征图和所述 I 帧特征图经由所述相加层得到在所述 I 帧特征图基础上经过相加操作后的 P 帧特征图；所述 I 帧特征图和所述经过相加操作后的 P 帧特征图经由所述第一拼接层得到按照帧的先后顺序进行拼接操作后的拼接特征图。本申请实施例提供的技术方案通过累加变换提供与 I 帧更紧密的信息关联，经过神经网络中的 Add（相加）计算，提供更加全面的待识别信息。

以 H264 为例，图 2 为本申请实施例提供的一种压缩视频中帧排列示意图，如图 2 所示，从第二个 P 帧开始，每一个 P 帧的 MV 和 RGBR 都依赖于前面一帧的 P 帧（如 P2 帧依赖于 P1 帧），为了使 P 帧与 I 帧的关联更加紧密，可以对 P 帧的 MV 和 RGBR 做累加的变换，获取到输入神经网络的 P 帧的 MV 和 RGBR 是相对于前面的 I 帧（如累加变换后，P2 帧变为依赖前面的 I 帧），而不是相对于前一个 P 帧。上述相加操作可以是参照 ResNet 中的残差相加的方式直接对 MV 特征图和 RGBR 特征图按每个元素（像素）与 I 帧的输出（经过第一 2D 残

差网络处理后的 I 帧特征图) 进行相加, 之后再按帧的先后顺序对 3 个特征图进行拼接操作。

在一些实施例中, 在所述第一 2D 残差网络的残差结构前包括特征变换层; 所述分组视频数据在进入所述残差结构前, 经由所述特征变换层得到经过向上特征变换和/或向下特征变换的分组视频数据。本申请实施例提供的技术方案, 在进入残差 (residual) 结构之前, 也即进行卷积操作前, 可以通过特征变换 (Feature Shift, FS) 对特征图进行部分特征的变换 (shift) 操作, 使得一个特征图中包含不同时间点的特征图中的部分特征, 这样在进行卷积操作时, 特征图中包含了时序信息, 可以在基本不增加计算量的前提下有能力处理时序信息的采集和融合, 丰富待识别特征图的信息, 提升识别准确度, 相比于相关技术中的基于 3D 卷积或光流信息的方案来说, 可有效降低计算量。

示例性的, 图 3 为本申请实施例提供的一种特征变换操作示意图。如图 3 所示, 以 MV 特征图为例, 为了便于说明, 图 3 中仅示出了针对 3 个 MV 特征图进行特征变换操作的过程, 3 个 MV 特征图分别对应不同的时间点, 假设将 MV 特征图划分为 4 个部分, 假设针对第 1 部分进行向下特征变换 (Down Shift), 针对第 2 部分进行向上特征变换 (Up Shift), 经过变换后的第 2 个 MV 特征图则同时包含了 3 个时间点的 MV 特征图的部分特征。在实际应用时, 划分规则, 以及向上特征变换和/或向下特征变换作用区域可根据实际情况进行设置。

在一些实施例中, 基于预设提取策略提取所述区间压缩视频中的 P 帧数据, 包括: 采用等间隔方式提取所述区间压缩视频中的预设数量的 P 帧数据; 其中, 在所述第一预设模型的训练阶段, 采用随机间隔方式提取区间压缩视频中的预设数量的 P 帧数据。本申请实施例提供的技术方案, 可以增强第一预设模型的鲁棒性。

图 4 为本申请实施例提供的另一种动作识别方法的流程示意图, 在上述各示例实施例基础上进行细化, 例如, 该方法包括:

步骤 401、基于预设分组规则对原始压缩视频数据进行区间划分, 得到区间压缩视频。

在一些实施例中, 原始压缩视频可采用预设的视频和图像增强策略进行增强处理, 可根据业务的数据需求以及驱动等配置情况来选择增强策略。在模型训练和模型应用时可采用相同的增强策略。

步骤 402、基于预设提取策略提取区间压缩视频中的 I 帧数据和 P 帧数据。

其中，P 帧数据包含 P 帧对应的运动矢量信息和色素变化残差信息。

步骤 403、对 P 帧数据进行累加变换，以使得变换后的 P 帧数据依赖于前向相邻的 I 帧，根据 I 帧数据和变换后的 P 帧数据确定分组视频数据。

步骤 404、将分组视频数据输入至第一预设模型中，并根据第一预设模型的输出结果确定包含动作的目标分组视频数据。

示例性的，图 5 为本申请实施例提供的一种基于压缩视频流的动作识别过程示意图，压缩视频被划分为 n 个片段，以其中的 S2 为例，将提取到的 I 帧数据、MV 数据和 RGBR 数据输入至第一预设模型中。第一预设模型中包含第一 2D 残差网络、相加层、第一拼接层和第一全连接层（Fully connected layer, FC layer），其中，第一 2D 残差网络为 2D Res18，其残差结构前包括特征变换层（FS）。在一些实施例中，在第一预设模型的训练阶段，可根据实际需求选择训练方式，包括损失函数（loss）等，例如损失函数可采用交叉熵，还可以采用其他辅助损失函数来提高模型效果。另外，可使用一些非启发式的优化算法来提高随机梯度下降的收敛速度以及优化性能。

图 6 为本申请实施例提供的一种第一 2D 残差网络结构示意图，如图 6 所示，2D Res18 是一个 18 层的残差神经网络，由 4 个阶段（stage）共 8 个使用 2D 卷积的残差块（block）组成，网络结构较浅，为尽可能使用卷积层充分提取特征，8 个 2D 残差块可均采用 3*3 的卷积进行，也即卷积核参数为 3*3。其中，卷积层一般指一个用于完成局部像素值的加权求和以及非线性激活的网络层，为了减少卷积层运算的通道数进而减少参数量，每个 2D 残差块都可使用瓶颈（bottleneck）的设计理念，即每个残差块都由 3 个卷积层组成（卷积核参数分别为 1*1、3*3 和 1*1），首层和尾层分别用于压缩和恢复图像通道。当然，第一 2D 残差网络结构以及各参数也可根据实际需求进行调整。

如图 5 所示，I 帧数据、MV 数据和 RGBR 数据在进入 2D Res18 残差结构（也即 2D 残差块）前，分别经由 FS 层得到经过向上特征变换和向下特征变换的 I 帧数据、MV 数据和 RGBR 数据，然后再分别经由残差结构得到得出 C*H*W 一致的特征图，经由相加层参照 ResNet 中的残差相加的方式直接对 MV 特征图和 RGBR 特征图按每个元素与 I 帧的输出进行相加，之后再按帧的先后顺序对 3 个特征图进行 concat 操作，得到拼接特征图，拼接特征图经由第一全连接层（FC）得到是否包含动作的分类结果。

图 7 为本申请实施例提供的一种动作标签示意图，图 5 中最后 2 个圆圈表

示输出是二分类，设置动作标签为二分类是为了保证提高包含动作的片段的召回，标签粒度为是否包含动作，设计为只有“是”或者“否”的二分类级别，即不关心包含的是哪个细分类动作。如图 7 所示，在模型训练过程中，可将训练用的短视频切割分为 6 个 segments，其中 segment S2 包含动作 A1，segment S4 包含动作 A2，A1 和 A2 为 2 个不同的动作，但这 2 个位置对应的标签是一样的，都设置为 1，也即不区分 A1 和 A2 动作的不同。因而，在实际应用时，第一预设模型输出的结果也同样为“是”或“否”，实现压缩视频中动作的粗粒度识别。

步骤 405、对目标分组视频数据进行解码，得到待识别分组视频数据。

步骤 406、将待识别分组视频数据输入至第二预设模型中，并根据第二预设模型的输出结果确定待识别分组视频数据中包含的动作类型。

本申请实施例提供的动作识别方法，先基于压缩视频进行动作识别，在无需解压视频的情况下，提取 I 帧、P 帧的 MV 和 RGBR 信息，并对 MV、RGBR 做变化以增加对 I 帧的信息依赖，实现了以较少计算力需求处理大量的时长不定的视频，且使用无需计算力要求的 FS 来增加模型对时序信息的提取，增强模型能力的情况下没有导致计算效率的降低，重新设计动作的标签粒度，以保证召回的目标，让模型处理简单的二分类问题，可以提高召回率，随后对初步筛选出来的包含动作的视频片段做精确识别，在保证识别精度的前提下有效减少计算量，提高动作识别效率。

在一些实施例中，所述对所述目标分组视频数据进行解码，得到待识别分组视频数据，包括：对所述目标分组视频数据进行解码，得到待识别分段视频图像；获取所述待识别分段视频图像中的频域信息，根据所述频域信息生成对应的频域图；将所述待识别分段视频图像和对应的频域图作为待识别分组视频数据。本申请实施例提供的技术方案，可以丰富图像信息，提高第二预设模型的准确度。

在一些实施例中，所述第二预设模型包括基于用于在线视频理解的高效卷积网络（Efficient Convolutional Network for Online Video, ECO）架构的模型。ECO 架构可理解为视频特征提取器，提供了一种视频特征获取的架构设计，里面包含了 2D 特征提取网络和 3D 特征提取网络，该架构可以在得到较好性能的同时提高速度，本申请实施例可以在 ECO 架构基础上进行改进和设计，得到第二预设模型。

在一些实施例中，所述第二预设模型包括第二拼接层、第二 2D 残差网络、

3D 残差网络、第三拼接层和第二全连接层；所述待识别分组视频数据被输入至第二预设模型中后，经由所述第二拼接层得到对所述待识别分段视频图像和对应的频域图进行拼接后的拼接图像数据；所述拼接图像数据经由所述第二 2D 残差网络得到 2D 特征图；将所述第二 2D 残差网络的中间层输出结果作为所述 3D 残差网络的输入，并经由所述 3D 残差网络得到 3D 特征图；所述 2D 特征图和所述 3D 特征图经由所述第三拼接层得到拼接后的待识别特征图；所述待识别特征图经由所述第二全连接层得到对应的动作类型标签。本申请实施例提供的技术方案，在 ECO 架构基础上，将第二 2D 残差网络的中间层输出结果作为 3D 残差网络的输入，实现网络结构的复用，提升模型速度。

示例性的，待识别分组视频数据首先进入模型的 2D 卷积部分提取每张图像的特征，然后进入 3D 卷积部分提取动作的时序信息，最后输出多分类的结果。其中，第二 2D 残差网络和 3D 残差网络的网络结构和相关参数可根据实际需求设置，在一些实施例中，第二 2D 残差网络为 2D Res50，3D 残差网络为 3D Res10。

示例性的，所述第二预设模型还包括第一池化层和第二池化层；所述 2D 特征图在被输入至所述第三拼接层之前，经由所述第一池化层得到对应的包含第一元素数量的一维 2D 特征向量；所述 3D 特征图在被输入至所述第三拼接层之前，经由所述第二池化层得到对应的包含第二元素数量的一维 3D 特征向量。相应的，所述 2D 特征图和所述 3D 特征图经由所述第三拼接层得到拼接后的待识别特征图，包括：所述一维 2D 特征向量和所述一维 3D 特征向量经由所述第三拼接层得到拼接后的待识别向量。本申请实施例提供的技术方案，特征图在经过卷积层进行特征提取以后仍然有较大的尺寸，此时若直接把特征图展平，可能会使得特征向量维度过高，因此，2D 和 3D 残差网络在完成一系列的特征提取操作以后，可使用池化操作把特征图直接归总为一维特征向量，以降低特征向量的维度。其中，第一池化层和第二池化层可根据实际需求设计，例如可以是全局平均池化（Global average Pooling, GAP）或最大池化等。第一元素数量和第二元素数量也可以自由设置，例如可根据特征图的通道数来设置。

示例性的，所述第一池化层包括多感受野池化层。感受野一般指特征值所覆盖的图像或视频范围，用来表示网络内部的不同神经元对原图像的感受范围的大小，或者说，卷积神经网络每一层输出的特征图上的像素点在原始图像上映射的区域大小。感受野的值越大表示其能接触到的原始图像范围就越大，也意味着它可能蕴含更为全局、语义层次更高的特征；相反，值越小则表示其所

包含的特征越趋向局部和细节。因此感受野的值可以用来大致判断每一层的抽象层次。采用多感受野池化层的好处在于，使得特征可以对不同尺度的目标敏感，拓宽了对动作类别的识别范围。多感受野实现方式可根据实际需求设置。

在一些实施例中，所述第一池化层包括一级局部池化层、二级全局池化层和向量合并层，所述一级局部池化层中包含至少两个尺寸不相同的池化核；所述 2D 特征图经由所述一级局部池化层得到对应的至少两组不同尺度的 2D 池化特征图；所述至少两组不同尺度的 2D 池化特征图经由所述二级全局池化层得到至少两组特征向量；所述至少两组特征向量经由所述向量合并层得到对应的包含第一元素数量的一维 2D 特征向量。本申请实施例提供的技术方案，采用两级多感受野池化来归总特征图，利用不同尺寸的池化核进行多尺度池化，使得归总的特征拥有不同的感受野，同时也使得特征图尺寸大大减少，提高识别效率，随后利用二级全局池化来对特征进行归总，并利用向量合并层得到一维 2D 特征向量。

图 8 为本申请实施例提供的又一种动作识别方法的流程示意图，在上述各示例实施例基础上进行细化，以短视频审核场景为例，图 9 为本申请实施例提供的基于短视频的动作识别应用场景示意图，如图 9 所示，用户上传短视频后，先提取压缩视频流信息，并利用预先训练构造的基于压缩视频的动作视频模型（第一预设模型）识别出包含动作的目标片段，其他片段因不包含动作而被大量筛除，随后，对目标片段进行解码并提取图片时域频域信息，利用预先训练构造的基于解码图片的动作识别模型（第二预设模型）进行基于图片序列的更细粒度的动作识别，得到每个目标片段对应的动作类型，进而判断是否为目标动作，若不是目标动作，则也被筛除。

示例性的，该方法可包括：

步骤 801、基于预设分组规则对原始压缩视频数据进行区间划分，得到区间压缩视频。

其中，原始压缩视频数据为用户上传的短视频的视频流数据，每个区间压缩视频的长度可以是 5 秒。

步骤 802、基于预设提取策略提取区间压缩视频中的 I 帧数据和 P 帧数据。

其中，P 帧数据包含 P 帧对应的 MV 和 RGBR，每秒获取 1 个 I 帧以及其后对应的 2 个 P 帧，也即一个 5 秒的片段可以获取到 5 个 I 帧、10 个 MV 和 10 个 RGBR。

步骤 803、对 P 帧数据进行累加变换，以使得变换后的 P 帧数据依赖于前向相邻的 I 帧，根据 I 帧数据和变换后的 P 帧数据确定分组视频数据。

步骤 804、将分组视频数据输入至第一预设模型中，并根据第一预设模型的输出结果确定包含动作的目标分组视频数据。

步骤 805、对目标分组视频数据进行解码，得到待识别分段视频图像。

示例性的，从目标分组视频数据中解码出图像（图片），按时间顺序排成序列，得到序列图片。在一些实施例中，为了减少待识别数据量，可以按照预设获取策略从序列图片中获取设定数量的图像，作为待识别分段视频图像。预设获取策略例如等间隔获取，设定数量例如为 15。

步骤 806、获取待识别分段视频图像中的频域信息，根据频域信息生成对应的频域图。

其中，频域信息的采集方式不做限定，可针对序列图片中的每个图片采集频域信息，并生成与序列图片一一对应的频域（Frequency Domain，FD）图。

步骤 807、将待识别分段视频图像和对应的频域图作为待识别分组视频数据，输入至第二预设模型中，并根据第二预设模型的输出结果确定待识别分组视频数据中包含的动作类型。

示例性的，一个片段中的序列图片和对应的频域图将对应多分类标签中的一个标签，也即，第二预设模型的输出结果中，每个片段对应一个标签，若一个片段中包含多个动作，则模型会将主要动作对应的类型作为该片段的标签。例如，5 秒钟的 A 片段中有 4 秒钟存在动作，3 秒钟的动作为挥手，1 秒钟的动作为踢腿，则该片段 A 对应的动作标签为挥手。

图 10 本申请实施例提供的一种基于序列图片的动作识别过程示意图，其中 n 的数值一般与图 5 中的 n 的数值不相同，以片段 S2 为例，经过解码（Decode）后，得到序列图片（Image），提取频域信息后，生成对应的频域图（FD）。第二预设模型可基于 ECO 架构设计，其中可包含第二拼接层、第二 2D 残差网络、3D 残差网络、第一池化层、第二池化层、第三拼接层和第二全连接层。在一些实施例中，在第二预设模型的训练阶段，可根据实际需求选择训练方式，包括损失函数等，例如损失函数可采用交叉熵，还可以采用其他辅助损失函数来提高模型效果。另外，可使用一些非启发式的优化算法来提高随机梯度下降的收敛速度以及优化性能。如图 10 所示，解码后的序列图片和对应的频域图经过第二预设模型中第二拼接层后得到拼接图像数据，在一些实施例中，随后拼接图

像数据还可经过卷积层 (conv) 和最大池化层 (maxpool) 的处理后, 再输入到第二 2D 残差网络 (2D Res50) 中。2D Res50 的输出经过第一池化层 (多感受野池化层) 后得到 1024 维 2D 特征向量 (可理解为包含 1024 个元素的一维的行向量或列向量)。2D Res50 的中间输出结果将作为 3D 残差网络 (3D Res10) 的输入, 经过第二池化层 (GAP) 后得到 512 维 3D 特征向量。1024 维 2D 特征向量和 512 维 3D 特征向量经过第三拼接层后, 得到待识别特征图, 也即待识别特征向量, 最后经由第二全连接层得到最终的动作类型标签。

图 11 为本申请实施例提供的一种第二 2D 残差网络结构示意图, 如图所示, 第二 2D 残差网络为 2D Res50, 是一个 50 层的残差神经网络, 由 4 个 stage 共 16 个使用 2D 卷积的残差块组成。为了减少卷积层运算的通道数进而减少参数量, 每个残差块都可使用 bottleneck 的设计理念, 即每个残差块都由 3 个卷积层组成, 其中进出口的两个 1×1 卷积分别用来压缩和还原特征图的通道数。另外, 因为每经过一个 stage, 需要把特征图的尺寸缩小至四分之一、通道扩大为两倍, 所以在每个 stage 的入口处都可使用 2D 投影残差块, 这种残差块在旁路中增加了一个 1×1 卷积层, 用来保证做逐像素相加操作时, 特征图的尺寸和通道数保持一致。同理, 只在每个 stage 的入口处使用 2D 投影残差块也可以减少网络参数。

图 12 为本申请实施例提供的一种 3D 残差网络结构示意图。每个片段取 N 个视频帧 (如前文所述的从序列图像中获取的 15 个图像与对应的频域图经过拼接操作后得到的帧图片) 通过 2D Res50 获得的中间层特征图组, 例如来自 stage2-block4 的特征图组可被组装为三维的视频张量, 张量形状为 (c,f,h,w), 其中, c 是帧图片的通道数, f 是视频帧数, h、w 分别指帧图片的高和宽, 视频张量被输入至 3D Res10, 以进行整个视频时空特征的提取。如图 12 所示, 3D Res10 仅由 3 个 stage 共 5 个残差块组成, 所有卷积层都使用三维的卷积核, 卷积过程中, 时间维度上的信息也将一起参与计算。同样为了减少网络参数, 3D Res10 可使用卷积层数更少的残差块, 并且可去除在 bottleneck 残差块中使用的通道数扩张技术。

示例性的, 3D Res10 在完成卷积操作后, 可通过全局平均池化来对时空范围内的像素值求平均值, 得到一个 512 维的视频时空特征向量, 全局平均池化的核尺寸例如可以是 $2 \times 7 \times 7$ 。而 2D Res50 可与一般的卷积神经网络不同, 并不使用简单的全局池化操作来归总特征图, 而可采用两级多感受野池化来归总特

征图。图 13 为本申请实施例提供的两级多感受野池化操作的运算过程示意图。如图 13 所示, 首先, 在第一级的局部池化操作中, 三个归总范围(池化核)不同的最大池化被用于对 2D 特征图上的像素进行归总(三个最大池化的核尺寸例如可以分别为 7×7 、 4×4 和 2×2), 不同大小的池化核使得归总的特征拥有不同的感受野, 使得特征可以对不同尺度的目标敏感, 拓宽了对目标类别的识别范围。经过了第一级的多尺度局部池化操作以后, 得到三组尺寸大大减少的 2D 池化特征图, 这些 2D 池化特征图包含了不同尺度的视频特征。然后, 三组特征图分别进行第二级的求和全局池化(global sum-pooling)操作, 每个通道上的特征图都通过求和像素值得到一个特征, 三组特征图被归总为三个包含不同尺度视频空间信息的特征向量。最后对三个特征向量执行逐像素相加操作, 获得一个 1024 维的 2D 特征向量(如图 13 中最底部的圆圈所示)。经过前述过程, 视频的 N(如前文所述 15) 个帧将获得 N 个 1024 维的 2D 特征向量, 对这些帧图片的特征向量求和取平均即可获得代表整个视频空间信息的 2D 特征向量(即图 10 所示的 1024 维 2D 特征向量)。

本申请实施例提供的动作识别方法, 在上述实施例基础上, 在基于序列图片的动作识别部分, 从解码之后的图像中提取出频域信息拼接到图像中, 增加特征网络提取到的信息丰富度, 采用了 ECO 网络结构, 提取 2D 和 3D 信息后, 针对信息进行不同方式的池化, 并融合为一维的待识别向量, 对网络进行更多标签的分类, 提供了精度更高的识别。因为基于压缩视频流的动作识别部分已经过滤掉大量的非目标动作视频, 因此输入到对计算力需求较高的基于序列图片的动作识别部分的视频量较小, 综合两部分来说, 在保证动作识别精度的情况下, 以非常小的计算力需求完成了所有视频的检测任务, 有效提高识别效率, 做到了很好地平衡计算力需求、召回率和识别精度。

图 14 为本申请实施例提供的一种动作识别装置的结构框图, 该装置可由软件和/或硬件实现, 一般可集成在计算机设备中, 可通过执行动作识别方法进行动作识别。如图 14 所示, 该装置包括:

视频分组模块 1401, 设置为对原始压缩视频数据进行分组处理, 得到分组视频数据;

目标分组视频确定模块 1402, 设置为将所述分组视频数据输入至第一预设模型中, 并根据所述第一预设模型的输出结果确定包含动作的目标分组视频数据;

视频解码模块 1403，设置为对所述目标分组视频数据进行解码，得到待识别分组视频数据；

动作类型识别模块 1404，设置为将所述待识别分组视频数据输入至第二预设模型中，并根据所述第二预设模型的输出结果确定所述待识别分组视频数据中包含的动作类型。

本申请实施例提供的动作识别装置，在对压缩视频进行解压前，先利用第一预设模型粗略筛选出包含动作的视频片段，再利用第二预设模型精确识别包含的动作的类型，可以在保证识别精度的前提下有效减少计算量，提高动作识别效率。

本申请实施例提供了一种计算机设备，该计算机设备中可集成本申请实施例提供的动作识别装置。图 15 为本申请实施例提供的一种计算机设备的结构框图。计算机设备 1500 包括存储器 1501、处理器 1502 及存储在存储器 1501 上并可在处理器 1502 上运行的计算机程序，所述处理器 1502 执行所述计算机程序时实现本申请实施例提供的动作识别方法。

本申请实施例还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行本申请实施例提供的动作识别方法。

上述实施例中提供的动作识别装置、设备以及存储介质可执行本申请任意实施例所提供的动作识别方法，具备执行该方法相应的功能模块。未在上述实施例中详尽描述的技术细节，可参见本申请任意实施例所提供的动作识别方法。

注意，上述仅为本申请的实例实施例。本领域技术人员会理解，本申请不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本申请的保护范围。因此，虽然通过以上实施例对本申请进行了较为详细的说明，但是本申请不仅仅限于以上实施例，在不脱离本申请构思的情况下，还可以包括更多其他等效实施例，而本申请的范围由权利要求范围决定。