

动作识别方法和装置

相关申请的交叉引用

本申请要求于 2021 年 4 月 9 日提交的申请号为 202110380638.2、发明名称为“动作识别方法和装置”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

本公开涉及计算机技术领域，具体涉及动作识别方法和装置。

背景技术

通过识别视频中的检测对象所发生的动作，有利于对视频进行分类或识别视频的特征等。相关技术中的识别视频中的检测对象所发生的动作的方法，是采用基于深度学习方法训练的识别模型识别视频中的动作，或者是基于视频画面出现的动作的特征及其与预设特征之间的相似度，识别视频中的动作。

发明内容

本公开提供了一种动作识别方法、装置、电子设备以及计算机可读存储介质。

本公开的一些实施例提供了一种动作识别方法，包括：获取视频片段，并确定视频片段中的至少两个目标对象；针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从多个时空图子集中确定出终选子集；将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别。

在一些实施例中，目标对象在视频片段的各个视频帧中的位置基于以

下方法确定：获取目标对象在视频片段的起始帧中的位置，将起始帧作为当前帧，并通过多轮迭代操作确定目标对象在各个视频帧中的位置；迭代操作包括：将当前帧输入预先训练完成的预测模型，以预测目标对象在当前帧的下一帧中的位置，响应于确定当前帧的下一帧不是视频片段的终止帧，将本轮迭代操作中的当前帧的下一帧作为下一轮迭代操作的当前帧；响应于确定当前帧的下一帧是视频片段的终止帧，停止迭代操作。

在一些实施例中，连接该目标对象在视频片段的各个视频帧中的位置，包括：将目标对象在各个视频帧中以矩形框的形式表示；将各个视频帧中的矩形框依照各个视频帧的播放顺序进行连接。

在一些实施例中，将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，包括：将至少两个时空图中、相邻的时空图划分为同一个时空图子集。

在一些实施例中，获取视频片段，包括：获取视频，并将视频截取为各个视频片段；方法包括：将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集。

在一些实施例中，从多个时空图子集中确定出终选子集，包括：从多个时空图子集中确定出多个目标子集；基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集。

在一些实施例中，方法包括：获取时空图子集中、每一个时空图的特征向量；获取时空图子集中、多个时空图之间的关系特征；从多个时空图子集中确定出多个目标子集，包括：基于时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征，并利用高斯混合模型对多个时空图子集进行聚类，以及确定出用于表征每一类时空图子集的至少一个目标子集。

在一些实施例中，获取时空图子集中、每一个时空图的特征向量，包括：采用卷积神经网络获取时空图的空间特征、以及视觉特征。

在一些实施例中，获取时空图子集中、多个时空图之间的关系特征，包括：针对多个时空图中的每两个时空图，根据该两个时空图的视觉特征，确定该两个时空图之间的相似度；根据该两个特征图的空间特征，确定该

两个时空图之间的位置变化特征。

在一些实施例中，基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集，包括：针对多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；将每一个时空图子集与该目标子集之间的相似度中、最大的相似度，确定为该目标子集的分值；将多个目标子集中分值最大的目标子集，确定为终选子集。

本公开的一些实施例提供了一种动作识别装置，包括：获取单元，被配置为获取视频片段，并确定视频片段中的至少两个目标对象；构建单元，被配置为针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；第一确定单元，被配置为将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从多个时空图子集中确定出终选子集；识别单元，被配置为将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的的动作的动作类别。

在一些实施例中，目标对象在视频片段的各个视频帧中的位置基于以下方法确定：获取目标对象在视频片段的起始帧中的位置，将起始帧作为当前帧，并通过多轮迭代操作确定目标对象在各个视频帧中的位置；迭代操作包括：将当前帧输入预先训练完成的预测模型，以预测目标对象在当前帧的下一帧中的位置，响应于确定当前帧的下一帧不是视频片段的终止帧，将本轮迭代操作中的当前帧的下一帧作为下一轮迭代操作的当前帧；响应于确定当前帧的下一帧是视频片段的终止帧，停止迭代操作。

在一些实施例中，构建单元，包括：构建模块，被配置为将目标对象在各个视频帧中以矩形框的形式表示；连接模块，被配置为将各个视频帧中的矩形框依照各个视频帧的播放顺序进行连接。

在一些实施例中，第一确定单元，包括：第一确定模块，被配置为将至少两个时空图中、相邻的时空图划分为同一个时空图子集。

在一些实施例中，获取单元，包括：第一获取模块，被配置为获取视频，并将视频截取为各个视频片段；装置包括：第二确定模块，被配置为将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集。

在一些实施例中，第一确定单元，包括：第一确定子单元，被配置为从多个时空图子集中确定出多个目标子集；第二确定单元，被配置为基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集。

在一些实施例中，动作识别装置包括：第二获取模块，被配置为获取时空图子集中、每一个时空图的特征向量；第三获取模块，被配置为获取时空图子集中、多个时空图之间的关系特征；第一确定单元，包括：聚类模块，被配置为基于时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征，并利用高斯混合模型对多个时空图子集进行聚类，以及确定出用于表征每一类时空图子集的至少一个目标子集。

在一些实施例中，第二获取模块，包括：卷积模块，被配置为采用卷积神经网络获取时空图的空间特征、以及视觉特征。

在一些实施例中，第三获取模块，包括：相似度计算模块，被配置为针对多个时空图中的每两个时空图，根据该两个时空图的视觉特征，确定该两个时空图之间的相似度；位置变化计算模块，被配置为根据该两个特征图的空间特征，确定该两个时空图之间的位置变化特征。

在一些实施例中，第二确定单元，包括：匹配模块，被配置为针对多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；评分模块，被配置为将每一个时空图子集与该目标子集之间的相似度中、最大的相似度，确定为该目标子集的分值；筛选模块，被配置为将多个目标子集中分值最大的目标子集，确定为终选子集。

本公开的实施例提供了一种电子设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如上提供的动作识别方法。

本公开的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，其中，程序被处理器执行时实现如上提供的动作识别方法。

应当理解，本部分所描述的内容并非旨在标识本公开的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

附图用于更好地理解本方案，不构成对本申请的限定。其中：

图 1 是本申请的实施例可以应用于其中的示例性系统架构图；

图 2 是根据本申请的动作识别方法的一个实施例的流程图；

图 3 是根据本申请的动作识别方法的一个实施例中时空图构建方法的示意图；

图 4 是根据本申请的动作识别方法的一个实施例中时空图子集划分方法的示意图；

图 5 是根据本申请的动作识别方法的另一个实施例的示意图；

图 6 是根据本申请的动作识别方法的另一个实施例中时空图子集划分方法的示意图；

图 7 是根据本申请的动作识别方法的又一个实施例的流程图；

图 8 是根据本申请的动作识别装置的一个实施例的结构示意图；

图 9 是用来实现本申请实施例的动作识别方法的电子设备的框图。

具体实施方式

以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

图 1 示出了可以应用本申请的动作识别方法或动作识别装置的实施例的示例性系统架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 101、102、103 通过网络 104 与服务器 105 交互，以接收或发送消息等。终端设备 101、102、103 上可以安装有各种客户端应用，例如图像获取类应用、视频获取类应用、图像识别类应用、视

频识别类应用、播放类应用、搜索类应用、金融类应用等。

终端设备 101、102、103 可以是具有显示屏并且支持接收服务器消息的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、电子播放器、膝上型便携计算机和台式计算机等等。

终端设备 101、102、103 可以是硬件，也可以是软件。当终端设备 101、102、103 为硬件时，可以是各种电子设备，当终端设备 101、102、103 为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块（例如用来提供分布式服务的多个软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

服务器 105 可以获取终端设备 101、102、103 发送的视频片段，并确定视频片段中的至少两个目标对象；针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；将构建的至少两个时空图划分为多个时空图子集，并在这多个时空图子集中确定出终选子集；将终选子集所包含的时空图之间的关系所指示的目标对象之间的动作类别确定为该视频片段所包含的动作的动作类别。

需要说明的是，本公开的实施例所提供的动作识别方法一般由服务器 105 执行，相应地，动作识别装置一般设置于服务器 105 中。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器的。

继续参考图 2，示出了根据本公开的动作识别方法的一个实施例的流程 200，包括以下步骤：

步骤 201，获取视频片段，并确定视频片段中的至少两个目标对象。

在本实施例中，动作识别方法的执行主体（例如图 1 所示的服务器 105）可以通过有线或者无线的方式获取视频片段，并确定出该视频片段中的至少两个目标对象。其中，目标对象可以是人、可以是动物、也可以任何可以存在于视频画面中的实体。

在本实施例中，可以采用训练完成的目标识别模型识别出视频片段中的各个目标对象。也可以采用将视频画面与预设图形对比匹配等方式，识别视频画面中出现的目标对象。

步骤 202, 针对至少两个目标对象中的每一个目标对象, 连接该目标对象在视频片段的各个视频帧中的位置, 构建该目标对象的时空图。

在本实施例中, 针对至少两个目标对象中的每一个目标对象, 可以将目标对象在视频片段的各个视频帧中的位置进行连线, 以构建该目标对象的时空图。其中, 时空图是指将目标对象在视频片段的各个视频帧中的位置进行连线后, 所形成的穿越视频帧的图形。

在一些可选地实施例中, 连接该目标对象在视频片段的各个视频帧中的位置, 包括: 将目标对象在各个视频帧中以矩形框的形式表示; 将各个视频帧中的矩形框依照各个视频帧的播放顺序进行连接。

在该可选地实施例中, 如图 3 的 3(a) 所示, 可以将目标对象在各个视频帧中均以矩形框(或者进行目标识别后所生成的候选框)的形式表示, 并依照视频帧的播放顺序, 将各个视频帧中的、表示该目标对象的矩形框依次进行连接, 以形成如图 3 的 3(b) 所示的该目标对象的时空图。其中, 图 3 的 3(a) 中包含四个矩形框, 分别代表目标对象: 视图左下角的平台 3011、马背 3012、刷子 3013、以及人物 3014, 其中代表人物的矩形框以虚线的形式表示, 仅仅为了和与之重叠的刷子的矩形框进行区分展示。图 3 的 3(b) 中的时空图 3021、时空图 3022、时空图 3023、时空图 3024 分别表示平台 3011 的时空图、马背 3012 的时空图、刷子 3013 的时空图以及人物 3014 的时空图。

在一些可选地实施例中, 可以将目标对象的中心点在各个视频帧中的位置、依照各个视频帧的播放顺序进行连接, 以形成该目标对象的时空图。

在一些可选地实施例中, 可以将目标对象在各个视频帧中均以一个预设的形状进行表示, 并依照视频帧的播放顺序, 将各个视频帧中的表示该目标对象的形状依次进行连接, 以形成该目标对象的时空图。

步骤 203, 将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集, 并从多个时空图子集中确定出终选子集。

在本实施例中, 将至少两个目标对象构建的至少两个时空图划分为多个时空图子集, 并从多个时空图子集中确定出终选子集。终选子集可以是多个时空图子集中包含时空图最多的子集; 终选子集可以是在计算每两个时空图子集之间的相似度时, 其他时空图子集均与该终选子集相似度大于

阈值的子集；终选子集还可以是所包含的时空图位于画面中心区域的时空图子集。

在一些可选地实施例中，从多个时空图子集中确定出终选子集，包括：从多个时空图子集中确定出多个目标子集；基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集。

在该可选地实施例中，可以先从多个时空图子集中确定出多个目标子集，计算多个时空图子集中的每一个时空图子集、与多个目标子集中的每一个目标子集进行相似度，并根据相似度计算结果，从多个目标子集中确定出终选子集。

具体地，可以首先从多个时空图子集中确定出多个目标子集，该多个目标子集是用于代表多个时空图子集的子集，该多个目标子集可以通过对多个时空图子集进行聚类运算后，得到的可以代表每一类时空图子集的至少一个目标子集。

针对每一个目标子集，可以将多个时空图子集中的每一个时空图子集与该目标子集进行匹配，可以将获得匹配的时空图子集最多的目标子集确定为终选子集。例如，存在目标子集 A、目标子集 B，以及时空图子集 1、时空图子集 2、时空图子集 3，并且预设时空图子集之间的相似度大于 80% 的情况下可以确定两个时空图子集为匹配。若，时空图子集 1 与目标子集 A 的相似度为 85%、时空图子集 1 与目标子集 B 的相似度为 20%、时空图子集 2 与目标子集 A 的相似度为 65%、时空图子集 2 与目标子集 B 的相似度为 95%、时空图子集 3 与目标子集 A 的相似度为 30%、时空图子集 3 与目标子集 B 的相似度为 90%，则，可以确定在所有时空图子集中，与目标子集 A 匹配的时空图子集的数量为 1 个、与目标子集 B 匹配的时空图的数量为 2 个。此时可以将目标子集 B 确定为终选子集。

该可选地实施例首先确定出目标子集，并基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集，可以提高确定终选子集的准确性。

步骤 204，将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别。

在本实施例中，由于时空图是用于表征目标对象在连续的视频帧中的空间位置，时空图子集中包含了各种可以组合的时空图之间的位置关系或者形态关系，因此，时空图子集可以用于表征目标对象之间的位姿关系。而终选子集是从多个时空图子集中选出的可以表征全局时空图子集的子集，因此，终选子集中包含的时空图之间的位置关系或者形态关系，可以用于表征全局目标对象之间的位姿关系，即，终选子集所包含的时空图之间的关系所指示的、目标对象之间的位姿关系所称为的动作类别，即可以作为该视频片段所包含的动作的动作类别。

本实施例提供的动作识别方法，获取视频片段，并确定视频片段中的至少两个目标对象；针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从多个时空图子集中确定出终选子集；将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别，可以利用时空图之间的关系表示目标对象之间的位姿关系，并将可以表征全局时空图子集的终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别，可以提高识别视频中动作的准确性。

可选地，目标对象在视频片段的各个视频帧中的位置基于以下方法确定：获取目标对象在视频片段的起始帧中的位置，将起始帧作为当前帧，并通过多轮迭代操作确定目标对象在各个视频帧中的位置；迭代操作包括：将当前帧输入预先训练完成的预测模型，预测目标对象在当前帧的下一帧中的位置，响应于确定当前帧的下一帧不是视频片段的终止帧，将本轮迭代操作中的当前帧的下一帧作为下一轮迭代操作的当前帧；响应于确定当前帧的下一帧是视频片段的终止帧，停止迭代操作。

在本实施例中，可以首先获取视频片段的起始帧，并获取目标对象在该起始帧中的位置，并将该起始帧作为当前帧，并通过多轮迭代操作确定出目标对象在该视频片段的各个帧中的位置，迭代操作包括：将当前帧输入预先训练完成的预测模型中，预测目标对象在当前帧的下一帧中的位置，若确定当前帧的下一帧不是该视频片段的终止帧，则将本轮迭代操作

中的当前帧的下一帧作为下一轮迭代操作的当前帧，以通过本轮迭代操作所预测的目标对象在对应的视频帧中的位置，继续预测目标对象在之后的视频帧中的位置。若确定当前帧的下一帧是该视频片段的终止帧，则此时目标对象在该视频片段的各个帧中的位置均已预测完成，可以停止迭代操作。

上述预测过程即为，已知目标对象在视频片段第一帧中的位置，通过预测模型，预测目标对象在第二帧中的位置，再根据获得的目标对象在第二帧中的位置、预测目标对象在第三帧中的位置，由此通过目标对象在前一帧的位置预测目标对象在下一帧中的位置，直到获得目标对象在该视频片段的全部视频帧中的位置。

具体地，若视频片段长度为 T 帧，首先，采用预先训练的神经网络模型（例如，Faster Region-Convolutional Neural Networks，快速区域卷积神经网络）检测视频片段的第一帧中人或者物体的候选框（即用于表征目标对象的矩形框），并保留前 M 个分数最高的候选框 $B_1 = \{b_1^m | m=1, \dots, M\}$ 。同理，基于第 t 帧的候选框集合 B_t ，预测模型为第 $t+1$ 帧生成候选框集合 B_{t+1} ，即，基于第 t 帧中的任一候选框 b_t^m ，根据第 t 帧和第 $t+1$ 帧相同位置的视觉特征来估计 b_t^m 在下一帧中的运动趋势。

之后，采用池化操作得到第 t 帧和第 $t+1$ 帧在相同位置（例如，第 m 个候选框的位置）的视觉特征 $\mathbf{F}_t^m$ 和 $\mathbf{F}_{t+1}^m$ 。

最后，采用紧凑的双线性池化（compact bilinear pooling, CBP）操作来捕获两个视觉特征之间成对的相关性，并模拟相邻帧之间的空间交互作用：

$$CBP(\mathbf{F}_t^m, \mathbf{F}_{t+1}^m) = \sum_{j=1}^N \sum_{k=1}^N \langle \phi(\mathbf{F}_{t,j}^m), \phi(\mathbf{F}_{t+1,k}^m) \rangle \quad (1)$$

其中， N 是局部描述子的个数， $\phi(\cdot)$ 是低维映射函数， $\langle \cdot \rangle$ 是二阶多项式核。最后，将 CBP 层的输出特征输入预先训练的回归模型/回归层，以获得回归层输出的基于 b_t^m 的运动趋势所预测的 b_{t+1}^m 。由此，通过估计每个候选框的运动趋势可以获得随后帧中的候选框集合，并将这些候选框连接成时空图。

本实施例基于目标对象在视频片段的起始帧中的位置预测目标对象在各个视频帧中的位置，而非采用已知的视频片段中的各个视频帧直接识

别目标对象的位置，可以避免由于目标对象之间的相互动作导致的目标对象在某个视频帧中被遮挡、所造成的识别结果不能真实的反映目标对象在该相互动作下实际所处的位置的问题，从而可以提高预测目标对象在视频帧中的位置的准确性。

可选地，将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，包括：将至少两个时空图中、相邻的时空图划分为同一个时空图子集。

在本实施例中，将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集的方法可以是：将该至少两个时空图中相邻的时空图划分为同一个时空图子集。

例如，如图4所示，可以采用节点表示图3的3(b)中的各个时空图，即，采用节点401表征时空图3021、采用节点402表征时空图3022、采用节点403表征时空图3023、采用节点404表征时空图3024。可以将相邻的时空图划分为同一个时空图子集，如可以是将节点401与节点402划分为同一个时空图子集，可以是将节点402与节点403划分为同一个时空图子集、可以是将节点401、节点402与节点403划分为同一个时空图子集、还可以是将节点401、节点402、节点403以及节点404划分为同一个时空图子集等等。

本实施例将相邻的时空图划分为同一个时空图子集，有利于将表征具有相互动作的关系的目标对象的时空图划分为同一时空图子集，所确定出的各个时空图子集可以全面的表征视频片段中目标对象存在的各个动作，有利于提高识别动作的准确性。

需要说明的是，为了可以显性化说明基于视频片段中目标对象的时空图，识别视频片段所包含的动作的动作类别的方法，以及便于清晰地表达方法的各个步骤，本公开采用节点的形式表征时空图。在本公开所述方法的实际应用中，可以不将时空图以节点的方式表示，而直接采用时空图执行各个步骤。

需要说明的是，本公开各实施例所述的将多个节点划分为一个子图即为将节点所表征的时空图划分为一个时空图子集；节点的节点特征是节点所表征的时空图的特征向量、节点之间连线的特征是节点所表征的时空图

之间的关系特征；至少一个节点所组成的子图是该至少一个节点所表征的时空图所组成的时空图子集。

继续参考图 5，示出了根据本公开的动作识别方法的另一个实施例的流程 500，包括以下步骤：

步骤 501，获取视频，并将视频截取为各个视频片段。

在本实施例中，动作识别方法的执行主体（例如图 1 所示的服务器 105）可以通过有线或者无线的方式获取完整视频，并通过视频分段方法或者视频片段截取方法从获取到的完整视频中截取出各个视频片段。

步骤 502，确定存在于各个视频片段中的至少两个目标对象。

在本实施例中，可以采用训练完成的目标识别模型识别出存在于各个视频片段中的各个目标对象。也可以采用将视频画面与预设图形对比匹配等方式，识别视频画面中出现的目标对象。

步骤 503，针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图。

步骤 504，将针对至少两个目标对象构建的至少两个时空图中、相邻的时空图划分为同一个时空图子集，和/或将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集，并从多个时空图子集中确定出多个目标子集。

在本实施例中，可以将针对至少两个目标对象构建的至少两个时空图中的、相邻的时空图划分为同一时空图子集，以及将相邻视频片段中的、同一个目标对象的时空图划分为同一个时空图子集。并从多个时空图子集中确定出多个目标子集。

例如，如图 6 的（a）所示，从完整的视频中提取视频片段 1、视频片段 2、以及视频片段 3，构建如图 6 的（b）所示的目标对象在各个视频片段中的时空图。目标对象 A（平台）在视频片段 1 中的构建的时空图为 601、在视频片段 2 中的构建的时空图为 605、在视频片段 3 中的构建的时空图为 609。目标对象 B（马背）在视频片段 1 中的构建的时空图为 602、在视频片段 2 中的构建的时空图为 606、在视频片段 3 中未被识别出。目标对象 C（刷子）在视频片段 1 中的构建的时空图为 603、在视频片段 2 中的构建的时空图为 607、在视频片段 3 中的构建的时空图为 610。目标对

象 D (人物) 在视频片段 1 中的构建的时空图为 604、在视频片段 2 中的构建的时空图为 608、在视频片段 3 中的构建的时空图为 611。视频片段 3 中出现了新的目标对象 (背景景观) 612。在该示例中, 每个时空图均为对应视频片段中序号相同的目标对象的时空图 (如, 视频片段 1 中, 图 6 的 (b) 中的时空图 601 是图 6 的 (a) 中目标对象 601 的时空图)

采用节点的形式表征上述各个时空图, 以构建如图 6 的 (c) 所示的视频的完整节点关系图, 其中每个节点表征与其序号相同的时空图 (如节点 601 表征时空图 601)。

如图 6 的 (c) 中, 可以将节点 601、节点 605、节点 606 划分为同一个子图, 可以将节点 603、节点 604、节点 607、节点 608 划分为同一个子图等等。

步骤 505, 基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度, 从多个目标子集中确定出终选子集。

步骤 506, 将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别。

本实施例中对步骤 503、步骤 505、步骤 506 的描述与步骤 202、步骤 204、步骤 205 的描述一致, 此处不再赘述。

本实施例提供的动作识别方法, 将获取到的完整视频划分为各个视频片段, 以及确定出存在于每个视频片段中的各个目标对象, 构建该目标对象属于每一个视频片段的时空图, 并将相邻的时空图划分为同一个时空图子集, 和/或将相邻视频片段中, 同一个目标对象的时空图划分为同一个时空子集, 并从多个时空图子集中确定出多个目标子集。由于同一视频片段的相邻的时空图体现了目标对象之间的位置关系, 相邻视频片段中同一目标对象的时空图可以体现该目标对象在视频播放过程中的位置的变化状态, 将同一视频片段中相邻的时空图, 和/或相邻视频片段中同一目标对象的时空图划分为同一时空图子集, 有利于将表征目标对象的动作变化的时空图划分为同一时空图子集, 所确定出的各个时空图子集可以全面的表征视频片段中目标对象存在的各个动作, 有利于提高识别动作的准确性。

继续参考图 7, 示出了根据本公开的动作识别方法的又一个实施例的

流程 700，包括以下步骤：

步骤 701，获取视频片段，并确定视频片段中的至少两个目标对象。

步骤 702，针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图。

步骤 703，将针对至少两个目标对象构建的多个时空图划分为多个时空图子集。

在本实施例中，将至少两个目标对象构建的至少两个时空图划分为多个时空图子集。

步骤 704，获取时空图子集中、每一个时空图的特征向量。

在本实施例中，可以获取时空图子集中，每一个时空图的特征向量。具体地，时空图所在的视频片段输入预先训练好的神经网络模型中，以获得该神经网络模型输出的每一个时空图的特征向量。该神经网络模型可以是循环神经网络、深度神经网络、深度残差神经网络等等。

在一些可选地实施例中，获取时空图子集中、每一个时空图的特征向量，包括：采用卷积神经网络获取时空图的空间特征、以及视觉特征。

在该可选地实施例中，时空图的特征向量包括时空图的空间特征、以及时空图的视觉特征。可以将时空图所在的视频片段输入预先训练完成的卷积神经网络中，以获得卷积神经网络输出的维度为 $T*W*H*D$ 的卷积特征，其中， T 代表卷积的时间维度、 W 代表卷积特征的宽度、 H 代表卷积特征的高度、 D 代表卷积特征的通道数。在该实施例中，为了保留原始视频的时间粒度，可以使卷积神经网络在时间维度上不存在下采样层，即不对视频片段的空间特征进行下采样。对于时空图在各帧中的边界框的空间坐标，对卷积神经网络输出的卷积特征执行池化操作，从而得到该时空图的视觉特征 f_v^{visual} 。将时空图在每一帧中的边界框的空间位置（例如，矩形框形状的时空图的中心点坐标以及矩形框的长、宽、高这一四维向量 $\hat{f}_v^{coord}$ ）输入多层感知机中，并将多层感知机的输出作为该时空图的空间特征 f_v^{coord} 。

步骤 705，获取时空图子集中、多个时空图之间的关系特征。

在本实施例中，可以获取时空图子集中，多个时空图之间的关系特征，其中，关系特征是表征特征之间的相似度、特征图之间的位置关系的特征。

在一些可选地实施例中，获取时空图子集中、多个时空图之间的关系

特征, 包括: 针对多个时空图中的每两个时空图, 根据该两个时空图的视觉特征, 确定该两个时空图之间的相似度; 根据该两个特征图的空间特征, 确定该两个时空图之间的位置变化特征。

在该可选地实施例中, 时空图之间的关系特征可以包括时空图之间的相似度或者时空图之间的位置变化特征, 针对多个时空图中的每两个时空图, 可以根据该两个时空图的视觉特征之间的相似度, 确定该两个时空图之间的相似度, 具体地, 可以通过如下公式 (2), 计算得到两个时空图之间的相似度:

$$f_{c_q}^{sem} = \varphi(f_{v_i}^{visual})^T \varphi(f_{v_j}^{visual}) \quad (2)$$

其中, $f_{c_q}^{sem}$ 代表时空图 v_i 和时空图 v_j 之间的相似度, $f_{v_i}^{visual}$ 和 $f_{v_j}^{visual}$ 分别代表时空图 v_i 和时空图 v_j 的视觉特征, $\varphi(\cdot)$ 代表特征转换函数。

在该可选地实施例中, 可以根据两个特征图的空间特征, 确定该两个时空图之间的位置变化信息, 具体地, 可以通过如下公式 (3), 计算得到两个时空图之间的位置变化信息:

$$\hat{f}_{c_q}^{coord} = \hat{f}_{v_i}^{coord} - \hat{f}_{v_j}^{coord} \quad (3)$$

其中, $\hat{f}_{c_q}^{coord}$ 代表时空图 v_i 和时空图 v_j 之间的位置变化信息, $\hat{f}_{v_i}^{coord}$ 以及 $\hat{f}_{v_j}^{coord}$ 分别代表时空图 v_i 和时空图 v_j 的空间特征。将该位置变化信息输入多层感知机后, 可以获得该多层感知机输出的时空图 v_i 和时空图 v_j 之间的位置变化特征 $f_{c_q}^{coord}$ 。

步骤 706, 基于时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征, 并利用高斯混合模型对多个时空图子集进行聚类, 以及确定出用于表征每一类时空图子集的至少一个目标子集。

在本实施例中, 可以基于时空图子集所包含的时空图的特征向量、以及时空图子集所包含的时空图之间的关系特征, 并利用高斯混合模型对多个时空图子集进行聚类, 以及确定出用于表征每一类时空图子集的每一个目标子集。

具体地, 可以将如图 6 的 (c) 所示的节点图分解为如图 6 的 (d) 所示的多个尺度子图, 不同尺度的子图中包含的节点数不同, 针对每一个尺度的子图, 可以将该子图所包含的各个节点的节点特征 (节点的节点特征即为其所表征的时空图的特征向量)、以及各个节点之间的连线特征 (两

个节点之间的连线特征即为两个节点所表征的两个时空图之间的关系特征)输入预设的高斯混合模型,利用高斯混合模型对该尺度的子图进行聚类,并确定出每一类子图中可以表征该类子图的目标子图。在利用高斯混合模型对同一尺度的子图进行聚类时,高斯混合模型输出的 k 个高斯核即为 k 个目标子图。

可以理解,目标子图中包含的节点所表征的时空图、组成了目标时空图子集。该目标时空图子集可以理解为能够代表这一尺度时空图子集的子集,该目标时空图子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别可以理解为该尺度下具有代表性的动作类别。由此, k 个目标子集可以视为与该尺度的子图对应的动作类别的标准模式。

步骤 707,基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度,从多个目标子集中确定出终选子集。

在本实施例中,可以基于多个时空图子集中的每一个时空图子集与多个目标子集中每一个目标子集之间的相似度,从多个目标子集中确定出终选子集。

具体地,针对图 6 的 (d) 所示的每一个子图,先通过如下公式获取该子图的混合权重:

$$\alpha = MLP(\mathbf{x}; \theta), \quad \hat{\gamma} = Softmax(\alpha), \quad (4)$$

其中,式中 $\mathbf{x}$ 代表子图 x 的特征,式中 $\mathbf{x}$ 包含子图 x 中各个节点的节点特征、以及节点之间连线的特征。 $\alpha = MLP(\mathbf{x}; \theta)$ 代表将 $\mathbf{x}$ 输入参数为 θ 的多层感知机,之后,将多层感知机的输出经过归一化指数函数 softmax 函数运算,并得到用于表征该子图的混合权重的 K 维向量 $\hat{\gamma}$ 。

通过上述公式 (4) 获得属于同一动作类别的 N 个子图的混合权重后,可以利用如下公式计算高斯混合模型中第 k ($1 \leq k \leq K$) 个高斯核的参数:

$$\hat{\phi}_k = \sum_{n=1}^N \frac{\hat{\gamma}_{nk}}{N} \quad (5)$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^N \hat{\gamma}_{nk} \mathbf{x}_n}{\sum_{n=1}^N \hat{\gamma}_{nk}} \quad (6)$$

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^N \hat{\gamma}_{nk} (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k) (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k)^\top}{\sum_{n=1}^N \hat{\gamma}_{nk}} \quad (7)$$

其中, $\hat{\phi}_k$, $\hat{\boldsymbol{\mu}}_k$, $\hat{\Sigma}_k$ 分别是第 k 个高斯核的权重、均值和协方差, $\hat{\gamma}_{nk}$ 代表第 n 个子图的混合权重在第 k 维度上的向量。在得到所有高斯核的参数之后, 任一子图 $\mathbf{x}$ 属于目标子集对应的动作类别的概率 $p(\mathbf{x})$ (即, 任一子图 $\mathbf{x}$ 与目标子集的相似度) 可以通过公式 (8) 计算:

$$p(\mathbf{x}) = \sum_{k=1}^K \hat{\phi}_k \frac{\exp(-\frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^\top \hat{\Sigma}_k^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k))}{\sqrt{|2\pi \hat{\Sigma}_k|}} \quad (8)$$

其中, $|\cdot|$ 代表矩阵的行列式。

在本实施例中, 可以将每个尺度上含有 N 各个子图的批量损失函数定义如下:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{n=1}^N \log p(x_n) + \lambda R(\hat{\Sigma}) \quad (9)$$

$$\text{其中, } R(\hat{\Sigma}) = \sum_{k=1}^K \sum_{i=1}^{D_k} \frac{1}{\hat{\Sigma}_{kii}} \quad (10)$$

其中, $p(x_n)$ 是子图 x_n 的预测概率, $R(\hat{\Sigma})$ 是协方差矩阵 $\hat{\Sigma}$ 的约束函数, 用于限制 $\hat{\Sigma}$ 的对角线上的值收敛到合理的解而不是 0。 λ 是用于平衡公式 (9) 前后两部分的权重参数, 可以基于需求进行设置 (如, 可以设置为 0.05)。由于高斯混合层中的每个操作都是可微分的, 因此可以将梯度从高斯混合层反向传播给特征提取网络, 从而以端到端的方式优化整个网络框架。

在本实施例中, 通过上述公式 (8) 获取到任一子图 $\mathbf{x}$ 属于各个动作类别的概率后, 针对每一个动作类别, 可以将属于该动作类别的子图的概率的平均值, 作为该动作类别的分数, 并将得分最高的动作类别作为视频所包含的动作的动作类别。

步骤 708, 将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别。

本实施例中对步骤 701、步骤 702、步骤 708 的描述与步骤 201、步骤 202、步骤 204 的描述一致，此处不再赘述。

本实施例提供的动作识别方法，基于各个时空图子集所包含的时空图的特征向量以及所包含的时空图之间的关系特征，并利用高斯混合模型对多个时空图子集进行聚类，可以在未知聚类类别的情况下，基于多时空图子集所包含的时空图的特征向量以及所包含的时空图之间的关系特征、所呈现出的正态分布曲线，对多个时空图子集进行聚类，可以提高聚类效率以及聚类准确性。

在上述结合图 7 描述的实施例的一些可选的实现方式中，针对多个目标子集中的每一个目标子集，基于每一个时空图子集与该目标子集之间的相似度，确定终选子集，包括：针对多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；将每一个时空图子集与该目标子集之间的相似度中、最大的相似度，确定为该目标子集的分值；将多个目标子集中分值最大的目标子集，确定为终选子集。

在本实施例中，针对多个目标子集中的每一个目标子集，可以获取每一个时空图子集与该目标子集之间的相似度，将所有相似度中最大的相似度作为该目标子集的得分，针对全部目标子集，将得分最高的目标子集确定为终选子集。

进一步参考图 8，作为对上述各图所示方法的实现，本公开提供了一种动作识别装置的一个实施例，该装置实施例与图 2、图 5 或图 7 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 8 所示，本实施例的动作识别装置 800，包括：获取单元 801、构建单元 802、第一确定单元 803、识别单元 804。获取单元，被配置为获取视频片段，并确定视频片段中的至少两个目标对象；构建单元，被配置为针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；第一确定单元，被配置为将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从多个时空图子集中确定出终选子集；识别单元，被配置为将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别。

在一些实施例中，目标对象在视频片段的各个视频帧中的位置基于以下方法确定：获取目标对象在视频片段的起始帧中的位置，将起始帧作为当前帧，并通过多轮迭代操作确定目标对象在各个视频帧中的位置；迭代操作包括：将当前帧输入预先训练完成的预测模型，以预测目标对象在当前帧的下一帧中的位置，响应于确定当前帧的下一帧不是视频片段的终止帧，将本轮迭代操作中的当前帧的下一帧作为下一轮迭代操作的当前帧；响应于确定当前帧的下一帧是视频片段的终止帧，停止迭代操作。

在一些实施例中，构建单元，包括：构建模块，被配置为将目标对象在各个视频帧中以矩形框的形式表示；连接模块，被配置为将各个视频帧中的矩形框依照各个视频帧的播放顺序进行连接。

在一些实施例中，第一确定单元，包括：第一确定模块，被配置为将至少两个时空图中、相邻的时空图划分为同一个时空图子集。

在一些实施例中，获取单元，包括：第一获取模块，被配置为获取视频，并将视频截取为各个视频片段；装置包括：第二确定模块，被配置为将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集。

在一些实施例中，第一确定单元，包括：第一确定子单元，被配置为从多个时空图子集中确定出多个目标子集；第二确定单元，被配置为基于多个时空图子集中的每一个时空图子集、与多个目标子集中每一个目标子集之间的相似度，从多个目标子集中确定出终选子集。

在一些实施例中，动作识别装置包括：第二获取模块，被配置为获取时空图子集中、每一个时空图的特征向量；第三获取模块，被配置为获取时空图子集中、多个时空图之间的关系特征；第一确定单元，包括：聚类模块，被配置为基于时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征，并利用高斯混合模型对多个时空图子集进行聚类，以及确定出用于表征每一类时空图子集的至少一个目标子集。

在一些实施例中，第二获取模块，包括：卷积模块，被配置为采用卷积神经网络获取时空图的空间特征、以及视觉特征。

在一些实施例中，第三获取模块，包括：相似度计算模块，被配置为针对多个时空图中的每两个时空图，根据该两个时空图的视觉特征，确定该两个时空图之间的相似度；位置变化计算模块，被配置为根据该两个特

征图的空间特征，确定该两个时空图之间的位置变化特征。

在一些实施例中，第二确定单元，包括：匹配模块，被配置为针对多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；评分模块，被配置为将每一个时空图子集与该目标子集之间的相似度中、最大的相似度，确定为该目标子集的分值；筛选模块，被配置为将多个目标子集中分值最大的目标子集，确定为终选子集。

上述装置 800 中的各单元与参考图 2、图 5 或图 7 描述的方法中的步骤相对应。由此上文针对动作识别方法描述的操作、特征及所能达到的技术效果同样适用于装置 800 及其中包含的单元，在此不再赘述。

根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

如图 9 所示，是根据本申请实施例的动作识别方法的电子设备 900 的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字助理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

如图 9 所示，该电子设备包括：一个或多个处理器 901、存储器 902，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置（诸如，耦合至接口的显示设备）上显示 GUI 的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作（例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统）。图 9 中以一个处理器 901 为例。

存储器 902 即为本申请所提供的非瞬时计算机可读存储介质。其中，该存储器存储有可由至少一个处理器执行的指令，以使该至少一个处理器

执行本申请所提供的动作识别方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的动作识别方法。

存储器 902 作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的动作识别方法对应的程序指令/模块（例如，附图 8 所示的获取单元 801、构建单元 802、第一确定单元 803、识别单元 804）。处理器 901 通过运行存储在存储器 902 中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的动作识别方法。

存储器 902 可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据用于提取视频片段的电子设备的使用所创建的数据等。此外，存储器 902 可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器 902 可选包括相对于处理器 901 远程设置的存储器，这些远程存储器可以通过网络连接至用于提取视频片段的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

动作识别方法的电子设备还可以包括：输入装置 903、输出装置 904 以及总线 905。处理器 901、存储器 902、输入装置 903 和输出装置 904 可以通过总线 905 或者其他方式连接，图 9 中以通过总线 905 连接为例。

输入装置 903 可接收输入的数字或字符信息，以及产生与用于提取视频片段的电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置 904 可以包括显示设备、辅助照明装置（例如，LED）和触觉反馈装置（例如，振动电机）等。该显示设备可以包括但不限于，液晶显示器（LCD）、发光二极管（LED）显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

此处描述的系统和技术各种实施方式可以在数字电子电路系统、集成电路系统、专用 ASIC（专用集成电路）、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多

个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

这些计算程序（也称作程序、软件、软件应用、或者代码）包括可编程处理器的机器指令，并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的，术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置（例如，磁盘、光盘、存储器、可编程逻辑装置（PLD）），包括，接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置（例如，CRT（阴极射线管）或者LCD（液晶显示器）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或者、触觉输入）来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（LAN）、广域网（WAN）和互联网。

计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此

并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

本公开提供的动作识别方法、装置，获取视频片段，并确定视频片段中的至少两个目标对象；针对至少两个目标对象中的每一个目标对象，连接该目标对象在视频片段的各个视频帧中的位置，构建该目标对象的时空图；将针对至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从多个时空图子集中确定出终选子集；将终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为视频片段所包含的动作的动作类别，可以提高识别视频中动作的准确性。

根据本申请的技术解决了现有的识别视频中的动作的方法存在识别不准确的问题。

应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

权利要求书

1、一种动作识别方法，包括：

获取视频片段，并确定所述视频片段中的至少两个目标对象；

针对所述至少两个目标对象中的每一个目标对象，连接该目标对象在所述视频片段的各个视频帧中的位置，构建该目标对象的时空图；

将针对所述至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从所述多个时空图子集中确定出终选子集；

将所述终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为所述视频片段所包含的的动作的动作类别。

2、根据权利要求 1 所述的方法，其中，所述目标对象在所述视频片段的各个视频帧中的位置基于以下方法确定：

获取所述目标对象在所述视频片段的起始帧中的位置，将所述起始帧作为当前帧，并通过多轮迭代操作确定所述目标对象在所述各个视频帧中的位置；

所述迭代操作包括：

将所述当前帧输入预先训练完成的预测模型，以预测所述目标对象在所述当前帧的下一帧中的位置，响应于确定所述当前帧的下一帧不是所述视频片段的终止帧，将本轮迭代操作中的所述当前帧的下一帧作为下一轮迭代操作的当前帧；

响应于确定所述当前帧的下一帧是所述视频片段的终止帧，停止所述迭代操作。

3、根据权利要求 1 所述的方法，其中，所述连接该目标对象在所述视频片段的各个视频帧中的位置，包括：

将所述目标对象在所述各个视频帧中以矩形框的形式表示；

将所述各个视频帧中的矩形框依照所述各个视频帧的播放顺序进行连接。

4、根据权利要求 1 所述的方法，其中，所述将针对所述至少两个目标对象构建的至少两个时空图划分为多个时空图子集，包括：

将所述至少两个时空图中、相邻的时空图划分为同一个时空图子集。

5、根据权利要求 1 所述的方法，其中，所述获取视频片段，包括：

获取视频，并将所述视频截取为各个视频片段；

所述方法包括：

将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集。

6、根据权利要求 1 所述的方法，其中，所述从所述多个时空图子集中确定出终选子集，包括：

从所述多个时空图子集中确定出多个目标子集；

基于所述多个时空图子集中的每一个时空图子集、与所述多个目标子集中每一个目标子集之间的相似度，从所述多个目标子集中确定出终选子集。

7、根据权利要求 6 所述的方法，其中，所述方法包括：

获取所述时空图子集中、每一个时空图的特征向量；

获取所述时空图子集中、多个时空图之间的关系特征；

其中，所述从所述多个时空图子集中确定出多个目标子集，包括：

基于所述时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征，并利用高斯混合模型对所述多个时空图子集进行聚类，以及确定出用于表征每一类时空图子集的至少一个目标子集。

8、根据权利要求 7 所述的方法，其中，所述获取所述时空图子集中、每一个时空图的特征向量，包括：

采用卷积神经网络获取所述时空图的空间特征、以及视觉特征。

9、根据权利要求 7 所述的方法，其中，所述获取所述时空图子集中、

多个时空图之间的关系特征，包括：

针对所述多个时空图中的每两个时空图，根据该两个时空图的视觉特征，确定该两个时空图之间的相似度；

根据该两个特征图的空间特征，确定该两个时空图之间的位置变化特征。

10、根据权利要求 6 所述的方法，其中，所述基于所述多个时空图子集中的每一个时空图子集、与所述多个目标子集中每一个目标子集之间的相似度，从所述多个目标子集中确定出终选子集，包括：

针对所述多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；

将每一个时空图子集与该目标子集之间的相似度中、最大的相似度，确定为该目标子集的分值；

将所述多个目标子集中分值最大的目标子集，确定为所述终选子集。

11、一种动作识别装置，包括：

获取单元，被配置为获取视频片段，并确定所述视频片段中的至少两个目标对象；

构建单元，被配置为针对所述至少两个目标对象中的每一个目标对象，连接该目标对象在所述视频片段的各个视频帧中的位置，构建该目标对象的时空图；

第一确定单元，被配置为将针对所述至少两个目标对象构建的至少两个时空图划分为多个时空图子集，并从所述多个时空图子集中确定出终选子集；

识别单元，被配置为将所述终选子集所包含的时空图之间的关系所指示的、目标对象之间的动作类别确定为所述视频片段所包含的动作的动作类别。

12、根据权利要求 11 所述的装置，其中，所述目标对象在所述视频片段的各个视频帧中的位置基于以下方法确定：

获取所述目标对象在所述视频片段的起始帧中的位置，将所述起始帧作为当前帧，并通过多轮迭代操作确定所述目标对象在所述各个视频帧中的位置；

所述迭代操作包括：

将所述当前帧输入预先训练完成的预测模型，以预测所述目标对象在所述当前帧的下一帧中的位置，响应于确定所述当前帧的下一帧不是所述视频片段的终止帧，将本轮迭代操作中的所述当前帧的下一帧作为下一轮迭代操作的当前帧；

响应于确定所述当前帧的下一帧是所述视频片段的终止帧，停止所述迭代操作。

13、根据权利要求 11 所述的装置，其中，所述构建单元，包括：

构建模块，被配置为将所述目标对象在所述各个视频帧中以矩形框的形式表示；

连接模块，被配置为将所述各个视频帧中的矩形框依照所述各个视频帧的播放顺序进行连接。

14、根据权利要求 10 所述的装置，其中，所述第一确定单元，包括：

第一确定模块，被配置为将所述至少两个时空图中、相邻的时空图划分为同一个时空图子集。

15、根据权利要求 10 所述的装置，其中，所述获取单元，包括：

第一获取模块，被配置为获取视频，并将所述视频截取为各个视频片段；

所述装置包括：

第二确定模块，被配置为将相邻视频片段中，同一个目标对象的时空图划分为同一个时空图子集。

16、根据权利要求 11 所述的装置，其中，所述第一确定单元，包括：

第一确定子单元，被配置为从所述多个时空图子集中确定出多个目标

子集；

第二确定单元，被配置为基于所述多个时空图子集中的每一个时空图子集、与所述多个目标子集中每一个目标子集之间的相似度，从所述多个目标子集中确定出终选子集。

17、根据权利要求 16 所述的装置，其中，所述装置包括：

第二获取模块，被配置为获取所述时空图子集中、每一个时空图的特征向量；

第三获取模块，被配置为获取所述时空图子集中、多个时空图之间的关系特征；

所述第一确定单元，包括：

聚类模块，被配置为基于所述时空图子集所包含的时空图的特征向量、以及所包含的时空图之间的关系特征，并利用高斯混合模型对所述多个时空图子集进行聚类，以及确定出用于表征每一类时空图子集的至少一个目标子集。

18、根据权利要求 17 所述的装置，其中，所述第二获取模块，包括：

卷积模块，被配置为采用卷积神经网络获取所述时空图的空间特征、以及视觉特征。

19、根据权利要求 17 所述的装置，其中，所述第三获取模块，包括：

相似度计算模块，被配置为针对所述多个时空图中的每两个时空图，根据该两个时空图的视觉特征，确定该两个时空图之间的相似度；

位置变化计算模块，被配置为根据该两个特征图的空间特征，确定该两个时空图之间的位置变化特征。

20、根据权利要求 16 所述的装置，其中，所述第二确定单元，包括：

匹配模块，被配置为针对所述多个目标子集中的每一个目标子集，获取每一个时空图子集与该目标子集之间的相似度；

评分模块，被配置为将每一个时空图子集与该目标子集之间的相似度

中、最大的相似度，确定为该目标子集的分值；

筛选模块，被配置为将所述多个目标子集中分值最大的目标子集，确定为所述终选子集。

21、一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求 1-10 中任一项所述的方法。

22、一种存储有计算机指令的非瞬时计算机可读存储介质，其中，所述计算机指令用于使所述计算机执行权利要求 1-10 中任一项所述的方法。