(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 104217214 A
(43)申请公布日 2014.12.17

(21)申请号 201410415114.2

(22)申请日 2014.08.21

(71)申请人 广东顺德中山大学卡内基梅隆大学国际联合研究院
地址 528300 广东省佛山市顺德区大良街道办广东顺德中山大学卡内基梅隆大学国际联合研究院
申请人 中山大学

(72)发明人 林倞 王可泽 李亚龙 王小龙

(74)专利代理机构 广州粤高专利商标代理有限公司 44102
代理人 林丽明

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/02(2006.01)

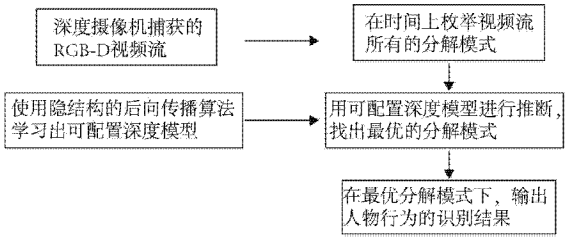
权利要求书3页 说明书6页 附图3页

(54)发明名称

基于可配置卷积神经网络的RGB-D人物行为识别方法

(57)摘要

本发明公开一种基于可配置卷积神经网络的RGB-D人物行为识别方法，构建基于可动态调整结构（可配置）的深度卷积神经网络；该识别方法可以直接处理RGB-D视频数据，并根据人物行为在时域上的变化动态调整网络结构，进而有效地自动抽取复杂人物行为的时空特征，最终大幅度提高人物行为识别的准确率。



CN 104217214 A

1. 一种基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，包括以下步骤：

S1. 构建可配置的深度模型，该深度模型引入隐变量，其构建过程为；

深度模型包括 M 个子网络和两个全连接层，每个子网络包括顺次连接的第一个三维卷积层、第一个降采样层、第二个三维卷积层、第二个降采样层和二维卷积层；M 个子网络的输出合并在一起，连接两个串联的全连接层；

在深度模型中引入隐变量，对输入的 RGB-D 视频帧在时间上进行划分，得到 M 个视频块，每个视频块作为一个子网络的输入；

S2. 学习深度模型的参数，通过隐式网络结构反向传播算法来学习深度模型的参数，其学习过程为：

固定当前深度模型参数进行人物行为识别，同时获取每个训练样本视频在时域上的优化分解模式；

固定输入视频的分解模式，使用反向传播算法学习网络的每层参数；

S3. 人物行为识别，在时间上枚举 RGB-D 视频流所有的分解模式，采用深度模型进行人物行为识别，获取最优分解模式，并在最优分解模式下输出人物行为的识别结果。

2. 根据权利要求 1 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，步骤 S1 中所述三维卷积层是指对输入 RGB-D 视频帧在时间域和空间域上同时做卷积，使用三维卷积层能够提取出人物的外观和运动信息；

设输入 RGB-D 视频帧的宽度和高度分别为 w 和 h，三维卷积核的大小为 w'×h'×m'，其中 w'，h'，m' 分别表示宽度，高度和时域上的长度，对从第 s 帧到 s+m'-1 帧的视频段应用三维卷积，能够获得一个特征图；

其中位于特征图 (x,y) 位置处的值表示成，

$$v_{xys} = tanh\left(b + \sum_{i=0}^{w'-1}\sum_{j=0}^{h'-1}\sum_{k=0}^{m'-1}\omega_{ijk} \cdot p_{(x+i)(y+j)(s+k)}\right) \qquad (1)$$

其中 $p_{(x+i)(y+j)(s+k)}$ 表示输入的第 (s+k) 帧中 (x+i,y+j) 位置的像素值，ω_{ijk} 表示卷积核的参数，b 表示跟与该特征图相关的偏置；

应用三维卷积得到 m-m'+1 个特征图，每个特征图的大小为 (w-w'+1,h-h'+1)，由于单个卷积核只能抽取一种类型的特征，则在每一层卷积层引入了多个卷积核抽取多种不同的特征，对于每一个子网络，分别将第一，第二个卷积层的卷积核数量定义为 c_1 和 c_2；

经过第一个三维卷积层操作后，得到了 c_1 个特征图集，每个包含 m-m'+1 个特征图；对于每一个特征图集，使用与第一个三维卷积相同的三维卷积的方法得到更高层级、新的特征图集；在 c_1 个特征图集上使用 c_2 个新的卷积核，在第二个三维卷积层得到 $c_1 \times c_2$ 个新的特征图集。

3. 根据权利要求 2 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，步骤 S1 中所述降采样层使用 max-pooling 操作，该操作是指对特征图按照最大值的策略进行降采样的过程，能够提取出保持形状和偏移不变性的特征；对于一组特征图，max-pooling 操作通过对它们降采样，得到同样数量的一组低分辨率特征图。

4. 根据权利要求 3 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特

征在于，步骤 S1 中所述二维卷积层是将三维卷积核的时域长度设置为 1，m' = 1，设二维卷积核的数量为 c_3，在已经得到的 $c_1 \times c_2$ 组特征图集上应用二维卷积核，最终得到 $c_1 \times c_2 \times c_3$ 组新的特征图集。

5. 根据权利要求 4 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，步骤 S1 中所述两层全连接层是在二维卷积层的输出上建立的感知机模型，两层全连接层分别为隐藏层和逻辑回归层；

将从 M 个子网络得到的特征图串联成一个长特征向量，该向量是从 RGB-D 视频中抽取到的特征；它的每一维元素都连向隐藏层的所有节点，并进一步全连接到网络顶部输出层所有的节点，共 K 个，等同于行为类别的数量 K；

每一个单元的输出看做输入视频中人的行为属于某类别的概率，为了归一化输出类别的概率，使用了 softmax 函数，即：

$$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \qquad (2)$$

z_i 是上一层的网络节点乘以第 i 个输出层的权重后的加权求和，$\sigma(z_i)$ 表示输出概率，且 $\sum_{i=1}^{K} \sigma(z_i) = 1$。

6. 根据权利要求 5 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，所述深度模型中每个子网络对应的输入的起始帧是可调整的，由隐变量控制；对于给定的输入 RGB-D 视频，使用前向传播算法来识别视频中人物的行为；

对于单个视频样本，定义 M 个子网络的起始帧点为 $(s_1, \ldots, s_M)$ 并且对应的输入帧的数量为 $(t_1, \ldots, t_M)$，其中 $1 \leqslant t_i \leqslant m$，

则深度模型的隐变量表示为 $H = (s_1, \ldots, s_M, t_1, \ldots, t_M)$，其表达的是每个子网络和视频段的对应关系；

给定输入视频 X，隐变量 H 以及模型的参数 ω，参数 ω 包括网络的边权重和偏置，识别的结果表达成向量 F(X, ω, H)，其中每个元素表示视频 X 属于某一行为类别的概率，将属于第 i 类的概率简记为 $F_i(X, \omega, H)$。

7. 根据权利要求 6 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，通过隐式网络结构反向传播算法来学习是过程为：

在学习时模型的参数 ω 和隐变量 H 必须同时进行优化，以两个步骤迭代地优化 ω 和 H 算法：

(101) 给定参数 ω，计算隐变量 H；

(102) 给定由隐变量 H 决定的输入帧，使用反向传播算法优化参数 ω；

假设共有 N 个训练样本 $(X_1, y_1), \ldots (X_N, y_N)$，其中 X_i 表示第 i 个输入视频 (i = 1, ……, N)，$y_i \in \{1, \ldots K\}$ 表示行为的类别，K 是类别的数量；

对所有样本定义了一组隐变量 $H = \{H_1, \ldots, H_N\}$，在训练过程中，使用逻辑回归定义损失函数 J(ω, H)，定义为，

$$J(\omega,\mathbf{H})=-\frac{1}{N}(\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}(y_i=k)\log F_k(X_i,\omega,H_i)+(1-\mathbf{1}(y_i=k))\log(1-F_k(X_i,\omega,H_i)))+\|\omega\|^2,$$

(3)

其中 1(•) ∈ {0,1} 是指示函数，损失函数的前两项表示似然的相反数，最后一项是正则项；

为了最小化损失 J(ω,H)，迭代地用以下步骤来优化参数 ω 和隐变量 H，

(201) 固定从上次迭代中优化的模型参数 ω，通过最大化对应于每个样本 (X_i, y_i) 的概率函数 $F_{y_i}(X_i,\omega,H_i)$ 来最小化公式 (3)，通过找到最优的隐变量 H 来实现，

$$H_i^*=argmax_{H_i}F_{y_i}(X_i,\omega,H_i). \qquad (4)$$

在输出结果上应用 softmax 分类，最大化 $F_{y_i}(X_i,\omega,H_i)$ 概率等价于降低样本属于其他类别的概率 $F_k(X_i,\omega,H_i)$，$\forall k\neq y_i$；

(202) 固定每个样本的隐变量，$H=\{H_1,\ldots,H_N\}$，得到输入 RGB-D 视频在时域上的分解模式，计算此时相应的损失 J(ω,H)，能够获得损失 J(ω,H) 相对于参数 ω 的梯度；通过应用反向传播算法，能够进一步地降低损失 J(ω,H) 同时优化网络模型参数 ω，

使用随机梯度下降算法更新模型的参数，并且每轮更新都使用所有的训练样本计算；

该优化算法在步骤 (201) 和 (202) 中迭代直到公式 (3) 收敛为止。

8. 根据权利要求 7 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，还包括基于海量普通视频的预训练，包括以下步骤：

1) 随机初始化网络参数；

2) 把每个普通视频从帧数上等分分解到子网络中；

3) 使用后向传播算法学习参数，将学习到的子网络的参数初始化深度模型，深度模型最终的输入是灰度和深度数据，将灰度通道的参数复制给深度信息的通道；

通过预学习来初始化子网络的参数，全连接层的参数是随机初始化。

9. 根据权利要求 8 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，步骤 S3 采用深度模型对 RGB-D 视频的人物行为进行识别，其具体过程为：

搜索类别行为标签 y 和隐变量 H 使概率 $F_i(X,\omega,H)$ 最大化，

$$(y^*,H^*)=\mathrm{argmax}_{(y,H)}F_y(X,\omega,H) \qquad (5)$$

通过优化隐变量 H 并计算出第 i 个样本属于每一个类别标签的概率 $F_y(X,\omega,H)$。选择最大概率；

对于 H 的领域空间 $H=(s_1,\ldots,s_M,t_1,\ldots,t_M)$，限制每个模型块所包含的输入帧数量为 $\tau\leqslant t_i\leqslant m$，并且不同的视频段不允许有重叠；

枚举在该限制条件下所有的 H 的取值情况，并通过前向算法求出概率 $F_y(X,\omega,H)$；通过选择最大的概率，得到更合适的 $F_y(X,\omega,H^*)$。

10. 根据权利要求 9 所述的基于可配置卷积神经网络的 RGB-D 人物行为识别方法，其特征在于，不同隐变量 H 决定的前向传播是相互独立的，能够通过并行计算来加速识别。

基于可配置卷积神经网络的 RGB-D 人物行为识别方法

技术领域

[0001] 本发明涉及人物行为识别领域，更具体地，涉及一种基于可配置卷积神经网络的 RGB-D 人物行为识别方法。

背景技术

[0002] 人物行为识别是计算机视觉研究的一个重要领域。它的应用包括智能监控、病人监护和一些涉及人机交互的系统。人物行为识别的目标是希望能够自动地从未知的视频中（例如，一段图像帧）分析和识别视频中正在发生的人物活动。简单来说，假如一个视频被分割成只包含一个单独的人物行为，系统的目标就是将该视频正确的分类到它所属的人物行为类别里。更一般的，人物行为识别希望能够持续地去识别视频中正在发生的人物活动，自动地标记出人物活动的开始时间和结束时间。

[0003] 人物行为识别是一个非常具有挑战的工作，识别的准确性很容易受到具体环境的影响。例如，以前的很多人物行为识别的工作使用的都是可见光摄像机拍摄的视频（或图像帧）数据，这些数据对人物的颜色、光线强度、遮挡以及复杂背景十分敏感，使得识别的准确率低。

[0004] 最近诞生的深度摄像机吸引了大批研究者的注意，且在视觉和机器人社区中有着广泛的应用。相对于传统的摄像机，深度摄像机提供了更丰富的场景信息（场景中物体距离摄像机的距离），并且能够在完全黑暗的环境中工作（这对一些病人监护系统、动物观测系统等有着很大的帮助）。深度摄像机捕获的视频称之为 RGB-D 视频。因而，深度摄像机的出现为人物姿势识别、动作行为识别等工作提供了更多的便利和可能。

[0005] 现有对 Kinect 深度摄像机获取的 RGB-D 视频中人物复杂行为活动的识别，这里存在着两个主要的难点：

[0006] (1) 对人物复杂行为的外观和运动信息的表达。由于人物个体的姿势和视角的不同，通常很难准确地抽取到人物的运动信息作为特征。同时，深度摄像机本身的机械噪声非常严重，使得人为的设计特征非常困难。

[0007] (2) 人物行为在时域上的变化太大。单个人物的行为可以看作是时间序列上发生的一系列子动作。例如，“用微波炉加热食物”可以被分解成拾取食物，走动和操作微波炉等几个子动作。如附图 2 所示，不同的人物在做相同的行为时，在时间上具有很大的差异（子动作持续的时间不同），使得识别非常困难。

[0008] 现有 RGB-D 人物行为识别的方法大多数是将视频表示成一系列固定长度的时间块，在该时间块上提取手工设计的特征，训练判别式或产生式的分类器来识别行为。由于手工设计的特征难以表达 RGB-D 视频数据中的运动信息，同时固定长度的时间块难以表达子动作在时间上的变化，其准确率不高。

发明内容

[0009] 为了克服现有技术的不足，本发明提出一种结合深度学习和动态结构调整的基

于可配置卷积神经网络的 RGB-D 人物行为识别方法，该人物行为识别方法可以直接处理 RGB-D 视频数据，有效地自动抽取复杂人物行为的时空特征，使得人物行为识别的准确率高。

[0010] 为了实现上述的目的，本发明的技术方案为：

[0011] 一种基于可配置卷积神经网络的 RGB-D 人物行为识别方法，包括：

[0012] S1. 构建可配置的深度模型，该深度模型包含隐变量，其构建过程为；

[0013] S11. 模型包括 M 个子网络和两个全连接层，每个子网络包括顺次连接的第一个三维卷积层、第一个降采样层、第二个三维卷积层、第二个降采样层和二维卷积层；M 个子网络的输出合并在一起，连接两个串联的全连接层；

[0014] S12. 在步骤 S11 的模型中引入隐变量，对输入的 RGB-D 视频帧在时间上进行划分，得到 M 个视频块，每个视频块作为一个子网络的输入；

[0015] S2. 深度模型的学习，通过隐式网络结构反向传播算法来学习，算法迭代为：

[0016] S21. 固定当前深度模型参数进行人物行为识别，同时获取每个训练样本视频在时域上的优化分解模式；

[0017] S22. 固定输入视频的分解模式，使用反向传播算法学习网络的每层参数；

[0018] S3. 采用深度模型对 RGB-D 视频的人物行为进行识别。

[0019] 与现有技术相比，本发明的有益效果为：

[0020] 本方法是将单个人物行为表示成一系列隐式的子动作，每个子动作都和一段不固定长度的类似立方体的视频段对应，利用深度网络，学习出一类人物行为在时域结构上特征，即可动态调整结构的深度卷积神经网络，具有以下特点：

[0021] 第一，深度结构是能自动从 RGB-D 数据中学习出有效的特征。首先，通过堆砌三维卷积层，降采样层以及全连接层构建出深度网络。其中，每个深度网络由 M 个子网络构成。每个子网络的输入是分割后的视频段。在子网络中，先应用两组 3D 卷积核和降采样操作，抽取相邻视频帧包含的运动信息，再应用 2D 卷积层抽取更抽象的高层语义信息；然后，将 M 个子网络的输出串联成一个长向量，使得每个视频段抽取的运动特征融合在一起，作为后两层全连接层的输入，最终得到行为的识别结果。

[0022] 第二，本发明公开的模型支持动态结构调整，是模型对复杂行为准确表达的关键。特别地，引入了隐变量来控制网络结构的动态调整。因此网络能够表达在时域上具有较大变化的人物行为。针对模型的特性，提出了一种两步迭代的优化方法来学习网络参数和确定隐变量，即隐结构的反向传播算法。

[0023] 采样本发明的方法能够解决了 RGB-D 视频中复杂人物行为识别所存在两个主要问题，可以直接处理 RGB-D 视频数据，进而有效地自动抽取复杂人物行为的时空特征，使得人物行为识别的准确率高。

附图说明

[0024] 图 1 是本发明系统的框图。

[0025] 图 2 是相同行为不同用户的展示图。

[0026] 图 3 是深度卷积神经网络示意图。

[0027] 图 4 是三维卷积示意图。

[0028] 图 5 是隐结构示意图。

[0029] 图 6 是隐结构的反向传播算法图。

具体实施方式

[0030] 下面结合附图对本发明做进一步的描述，但本发明的实施方式并不限于此。

[0031] 1. 结构化的深度模型

[0032] 首先详细介绍结构化深度模型及引入的隐变量。

[0033] 1.1 深度卷积神经网络

[0034] 为了对复杂的人物行为进行建模，在本实施方式中的深度模型如附图 3 所示。它由 M 个子网络和两个全连接层构成。其中，M 个子网络的输出串联成一个长向量，再接两个全连接层。（图 3 中 M 为 3，每个子网络用不同的图案来表示）每个子网络处理其相对应的视频段，该视频段跟一个从复杂行为中分解的子行为相关。每个子网络依次由三维卷积层、降采样层、三维卷积层、降采样层和二维卷积层级联构成。其中，三维卷积层能抽取出 RGB-D 视频的运动特征。降采样层能够对人物局部身体的变形进行很好的表达，同时对图像中的噪声不敏感。接下来详细的定义模型的各个重要部分。

[0035] 三维卷积层：三维卷积是指对输入 RGB-D 视频帧在时间域和空间域上同时做卷积，使用它能够提取出人物的外观和运动信息。假设输入 RGB-D 视频帧的宽度和高度分别为 w 和 h，三维卷积核的大小为 w' ×h' ×m'，其中 w',h',m' 分别表示宽度，高度和时域上的长度。如附图 4 所示，通过对从第 s 帧到 s+m'-1 帧的视频段应用三维卷积，可以获得一个特征图。其中位于特征图 (x,y) 位置处的值可以表示成，

[0036] $$v_{xys}=tanh\left(b+\sum_{i=0}^{w'-1}\sum_{j=0}^{h'-1}\sum_{k=0}^{m'-1}\omega_{ijk}\cdot p_{(x+i)(y+j)(s+k)}\right) \qquad (1.1)$$

[0037] 其中 $p_{(x+i)(y+j)(s+k)}$ 表示输入的第 (s+k) 帧中 (x+i,y+j) 位置的像素值，ω_{ijk} 表示卷积核的参数，b 表示跟与该特征图相关的偏置。故此可以得到 m-m'+1 个特征图，每个特征图的大小为 (w-w'+1,h-h'+1)。由于单个卷积核只能抽取一种类型的特征，因此在每一层卷积层引入了多个卷积核抽取多种不同的特征。对于每一个子网络，分别将第一，第二个卷积层的卷积核数量定义为 c_1 和 c_2。

[0038] 经过第一个三维卷积层操作后，得到了 c_1 个特征图集，每个包含 m-m'+1 个特征图。对于每一个特征图集，使用类似的三维卷积的方法得到更高层级、新的特征图集。由于在 c_1 个特征集上使用了 c_2 个新的第二三维卷积核，因而可以在下一层得到 $c_1\times c_2$ 个新的特征图集。

[0039] 降采样层：在本实施方式中降采样使用 max-pooling 操作。该操作是指对特征图按照一定策略（选取最大值）进行降采样的过程。这是一种被广泛应用的有效过程，它能够提取出保持形状和偏移不变性的特征。对于一组特征图，max-pooling 操作通过对它们降采样，得到同样数量的一组低分辨率特征图。更多地，如果在 $a_1\times a_2$ 大小的特征图上应用 2×2 的 max-pooling 操作，抽取 2×2 不重叠区域上的最大值，将得到大小为 $a_1/2\times a_2/2$ 的新特征图。

[0040] 二维卷积层：二维卷积可以看成是三维卷积的特例，即将三维卷积核的时域维度的长度设置为 1，例如，m' = 1。通过在一组特征图上应用二维卷积，可以得到同样数量的

一组新特征图。经过两层的二维卷积层以及 max-pooling 操作后，每组特征图在时间维度上都已经减小到足够小。在此基础上，继续应用二维卷积核来抽取特征图上更高层次的复杂特征。假设二维卷积核的数量为 c_3，并且在已经得到的 $c_1 \times c_2$ 组特征图集上应用这些二维卷积，最终得到 $c_1 \times c_2 \times c_3$ 组新的特征图集。

[0041] 全连接层：在模型中添加了两层全连接层，可以看做是在前面二维卷积层输出的基础上建立的感知机模型，全连接层分别隐藏层和逻辑回归层。首先将从 M 个子网络得到的特征图串联成一个长特征向量。该向量即是从 RGB-D 视频中抽取到的特征。它的每一维元素都连向第一个全连接层（隐藏层）的所有节点，并进一步全连接到所有的输出单元。输出单元共 K 个，等同于行为类别的数量 K，每一个单元的输出可以看做输入视频中人的行为属于某类别的概率。为了归一化输出类别的概率，使用了 softmax 函数，即

[0042] $$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \qquad (1.2)$$

[0043] z_i 是倒数第二层神经元乘以第 i 个输出层的权重后的加权求和。$\sigma(z_i)$ 表示输出概率，且 $\sum_{i=1}^{K} \sigma(z_i) = 1$。

[0044] 输入数据细节：首先从每个 RGB-D 视频中抽取出视频帧对应的灰度图和深度图。用两个通道分别存放灰度图和深度图。在进行卷积时，分别对这两个通道内应用三维卷积，并且将两个通道的卷积结果加在一起得到最终的卷积结果，这样使得卷积的特征图保持维度的一致。当然，模型可以应用到有更多通道的视频帧（例如进一步得到视频帧的梯度或光流等通道信息）。

[0045] 1.2 引入了隐变量的网络结构

[0046] 本实施方式的主要内容在于在深度模型结构中包含了隐变量。对于不同的包含人物行为的视频，每个子网络所对应的输入帧的起始点以及输入帧的帧数由隐变量控制。为了说明它，在附图 5 中展示了一个简单的例子，其中 3 个立方体块分别用不同的图案表示。对应起来讲，首先整个行为被分解成 3 个动作段，对应整个网络模型的 3 个子网络。每个子网络对应的输入的起始帧是可调整的，由隐变量控制。如果出现某些子网络所对应的输入帧的帧数不足 m 帧，那么子网络内部的部分单元将不会被激活（附图 5 中第一个和第三个子网络中黑色的点状圆圈）。对于给定的输入 RGB-D 视频，使用前向传播算法来识别视频中人物的行为。

[0047] 对于单个视频样本，定义 M 个子网络的起始帧点为 $(s_1, \ldots, s_M)$ 并且对应的输入帧的数量为 $(t_1, \ldots t_M)$，其中 $1 \leqslant t_i \leqslant m$。然后，模型的隐变量表示为 $H = (s_1, \ldots, s_M, t_1, \ldots, t_M)$，其表达的是每个子网络和视频段的对应关系。给定输入视频 X，隐变量 H 以及模型的参数 ω（包括网络的边权重和偏置），识别的结果可以表达成向量 $F(X, \omega, H)$，其中每个元素表示视频 X 属于某一行为类别的概率。并且，将其属于第 i 类的概率简记为 $F_i(X, \omega, H)$。

[0048] 2. 模型的学习——隐结构的反向传播算法

[0049] 由于在本实施方式的深度模型引入了隐变量，标准的反向传播算法不能优化模型的参数。因而，提出了一种隐结构的后向传播算法来学习模型参数。

[0050] 隐变量指示如何对输入视频在时域上进行划分。针对在学习时模型的参数 ω 和

隐变量 H 必须同时进行优化，提出了一种以下步骤迭代地优化 ω 和 H 算法：(i) 给定模型参数 ω，计算隐变量 H，如附图 6a；(ii) 给定由 H 决定的输入帧，使用反向传播算法优化模型参数 ω，如附图 6b。

[0051] 假设共有 N 个训练样本 $(X_1, y_1), \ldots (X_N, y_N)$，其中 X_i 表示输入视频，$y_i \in \{1, \ldots K\}$ 表示行为的类别并且 K 是类别的数量，$i = 1, \cdots\cdots, N$。为了更好的表达，同时对所有样本定义了一组隐变量 $H = \{H_1, \ldots, H_N\}$。在训练过程中，使用逻辑回归定义损失函数 $J(\omega, H)$，定义为，

[0052]

$$J(\omega,\mathbf{H}) = -\frac{1}{N}\left(\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}(y_i = k)\log F_k(X_i,\omega,H_i) + (1-\mathbf{1}(y_i = k))\log(1-F_k(X_i,\omega,H_i))\right) + \|\omega\|^2, \tag{2.1}$$

[0053] 其中 $1(\cdot) \in \{0, 1\}$ 是指示函数。损失函数 (2.1) 的前两项表示似然的相反数，而最后一项是正则项。

[0054] 为了最小化损失 $J(\omega, H)$，迭代地用以下步骤来优化参数 ω 和隐变量 H。

[0055] (a) 固定从上次迭代中优化的模型参数 ω，可以通过最大化对应于每个样本 (X_i, y_i) 的概率函数 $F_{y_i}(X_i, \omega, H_i)$ 来最小化公式 (2.1)，这可以通过找到最优的隐变量 H 来实现，

[0056] $$H_i^* = argmax_{H_i} F_{y_i}(X_i, \omega, H_i). \tag{2.2}$$

[0057] 需要提及的是，在输出结果上应用 softmax 分类，见公式 (1.2) 所示。最大化 $F_{y_i}(X_i, \omega, H_i)$ 概率等价于降低样本属于其他类别的概率 $F_k(X_i, \omega, H_i)$，$\forall k \neq y_i$。这样使得 log 似然增加从而降低损失 $J(\omega, H)$。

[0058] (b) 固定每个样本的隐变量，$H = \{H_1, \ldots, H_N\}$，可以得到输入 RGB-D 视频在时域上的分解模式。计算此时相应的损失 $J(\omega, H)$，能够获得 $J(\omega, H)$ 相对于参数 ω 的梯度。通过应用反向传播算法，能够进一步地降低损失 $J(\omega, H)$ 同时优化网络模型参数 ω。值得注意的是，使用随机梯度下降算法更新模型的参数，并且每轮更新都使用所有的训练样本来计算。

[0059] 该优化算法在步骤 (a)、(b) 两步中迭代直到公式 (2.1) 收敛为止。

[0060] 3. 模型的学习——基于海量普通视频的预训练

[0061] 对庞大的深度卷积神经网络参数来说，RGB-D 视频的数据量是太少了。为解决这一问题，同时提高识别的准确率。在本实施方式中采用了一种预训练的机制——使用传统的普通视频数据集来预训练。能够有监督地使用海量的、有动作类别标签的普通视频数据集来预训练模型。步骤如下：1) 随机初始化网络参数；2) 把每个普通视频从帧数上等分分解到子网络中；3) 使用传统的后向传播算法来学习参数，再将学习到的子网络的参数来初始化深度模型。值得注意的是，由于预训练是普通视频数据集，只学出了第一层的三维卷积核的灰度通道的参数，没有学到深度通道的参数。而最终的输入是灰度 / 深度数据，所以需要将灰度通道的参数复制给深度信息的通道 (D)。另外，由于高层语义需要从 RGB-D 数据集中学习，仅仅通过预学习来初始化子网络的参数，而全连接层的参数仍然是随机初始化。

[0062] 将整个的学习过程总结为算法 1。

[0063]

算法 1 学习框架

输入：

带类别标签的普通视频 和 RGB-D 人物行为数据集。

输出：

模型的参数 ω。

初始化：

使用普通视频人物行为数据集预训练时空卷积神经网络。

在 3D 人物行为数据上学习：

重复

1.固定模型参数，根据公式（2.2）来估计隐变量 **H**。

2.固定 **H**，根据公式（2.1）优化参数 ω。

直到 方程(2.1)中的损失 $J(\omega,\mathbf{H})$ 收敛。

[0064] 4. 人物行为识别

[0065] 人物行为识别是识别输入视频 X 所包含的人物行为。正式的，搜索类别行为标签 y 和隐变量 H 使概率 $F_i(X, \omega, H)$ 最大化，

[0066] $(y^*, H^*) = \mathrm{argmax}_{(y,H)} F_y(X, \omega, H)$ (3.1)

[0067] 通过优化 H 并搜索所有类别标签 y($1 \leqslant y \leqslant K$) 来计算最大的概率 $F_y(X, \omega, H)$。对于 H 的领域空间 $H = (s_1, \ldots, s_M, t_1, \ldots, t_M)$，限制每个模型块所包含的输入帧数量为 $\tau \leqslant t_i \leqslant m$，并且不同的视频段不允许有重叠（例如，$s_i+t_i \leqslant s_{i+1}$）。在本实施方式中，将 τ 设置成常数 $\tau = 4$。枚举在该限制条件下（该限制条件是指不同的视频段不允许有重叠）所有的 H 的取值情况，并通过前向算法求出概率 $F_y(X, \omega, H)$。通过选择最大的概率，得到更合适的 $F_y(X, \omega, H^*)$。由于不同 H 决定的前向传播是相互独立的，可以通过并行计算来加速识别。在本实施方式中，使用型号为英伟达 GTX TITAN 的显卡，处理一个 35 帧的视频，只需要 0.4 秒。

[0068] 以上所述的本发明的实施方式，并不构成对本发明保护范围的限定。任何在本发明的精神原则之内所作出的修改、等同替换和改进等，均应包含在本发明的权利要求保护范围之内。

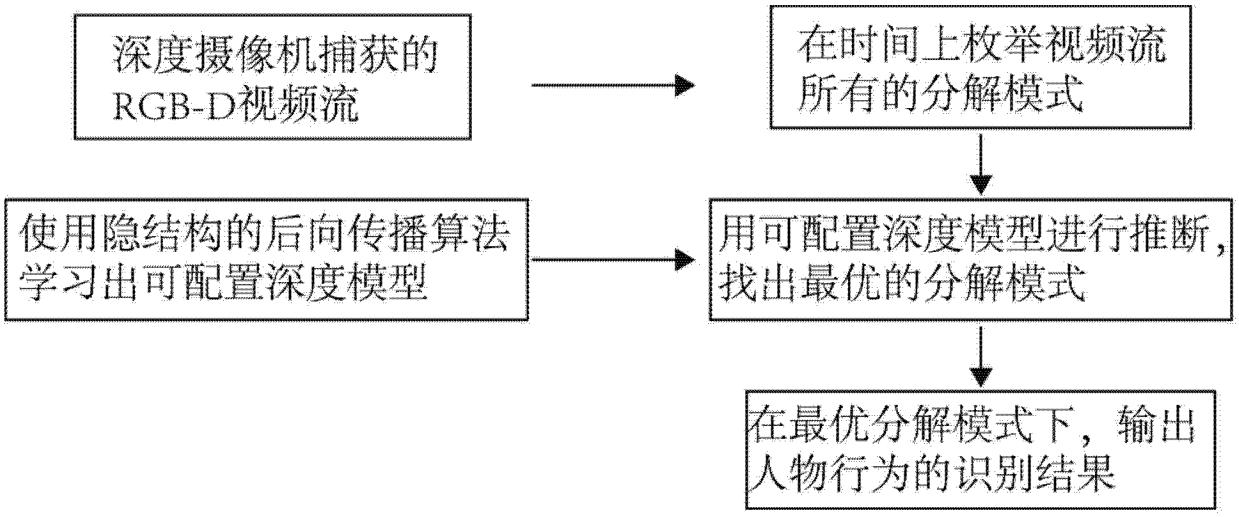

深度摄像机捕获的RGB-D视频流
在时间上枚举视频流所有的分解模式
使用隐结构的后向传播算法学习出可配置深度模型
用可配置深度模型进行推断，找出最优的分解模式
在最优分解模式下，输出人物行为的识别结果

图 1

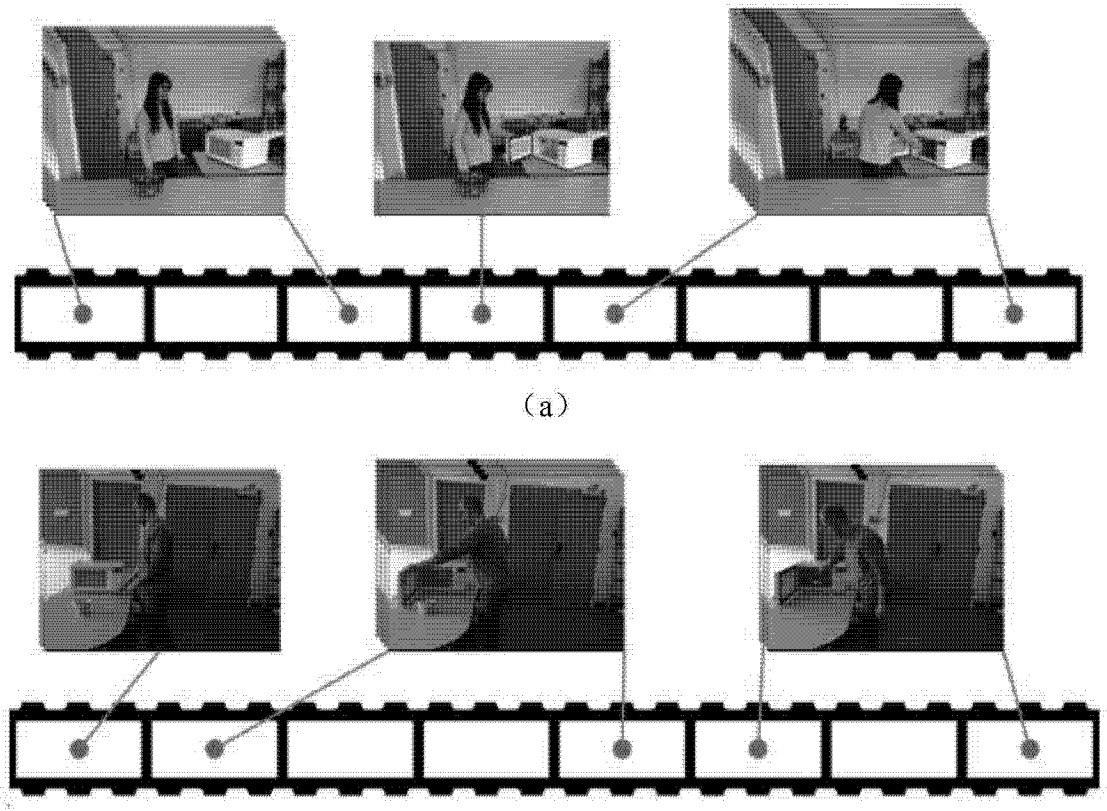

(a)
(b)

图 2

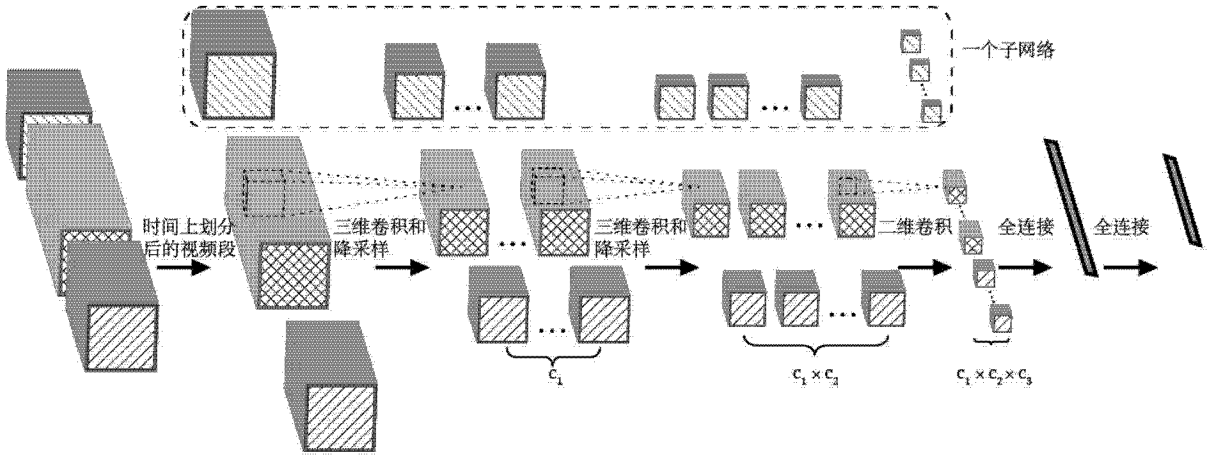

一个子网络
时间上划分
后的视频段
三维卷积和
降采样
三维卷积和
降采样
二维卷积
全连接
全连接
c_1
$c_1 \times c_2$
$c_1 \times c_2 \times c_3$

图 3

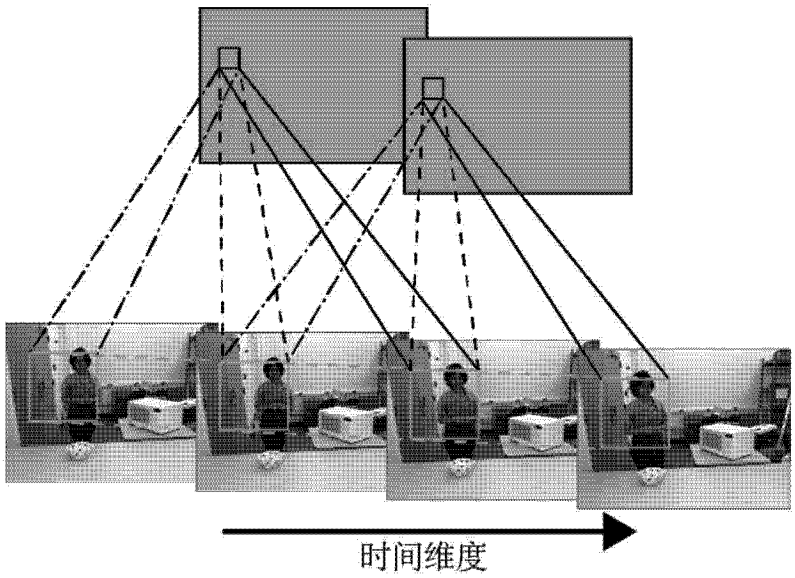

时间维度

图 4

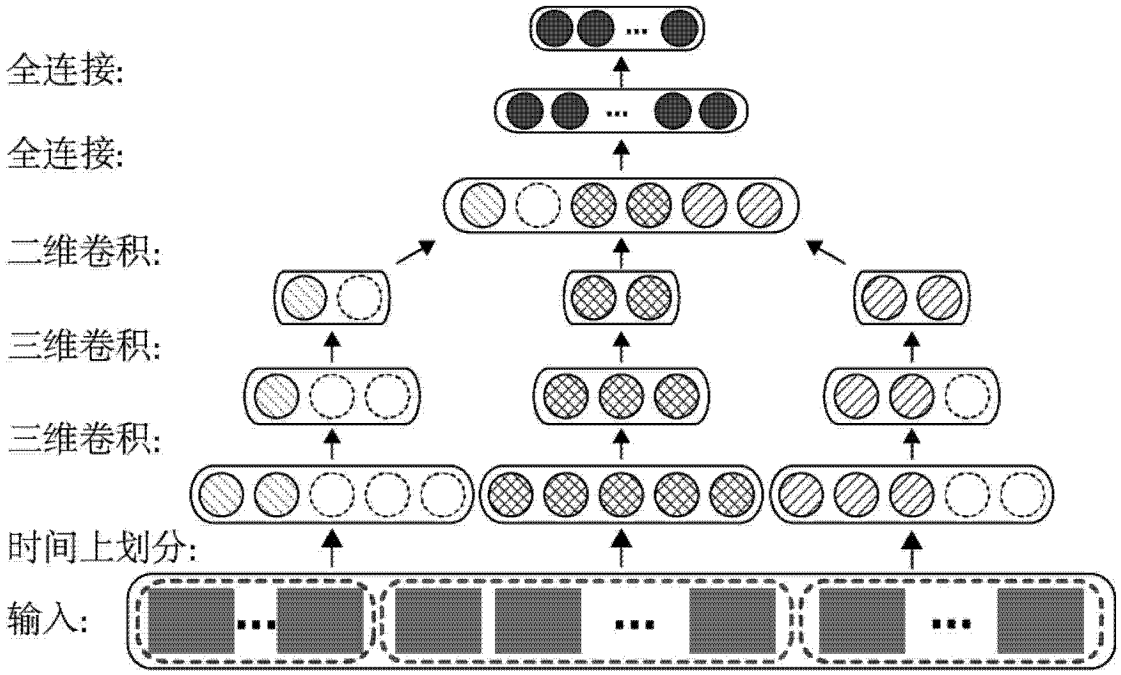


图 5

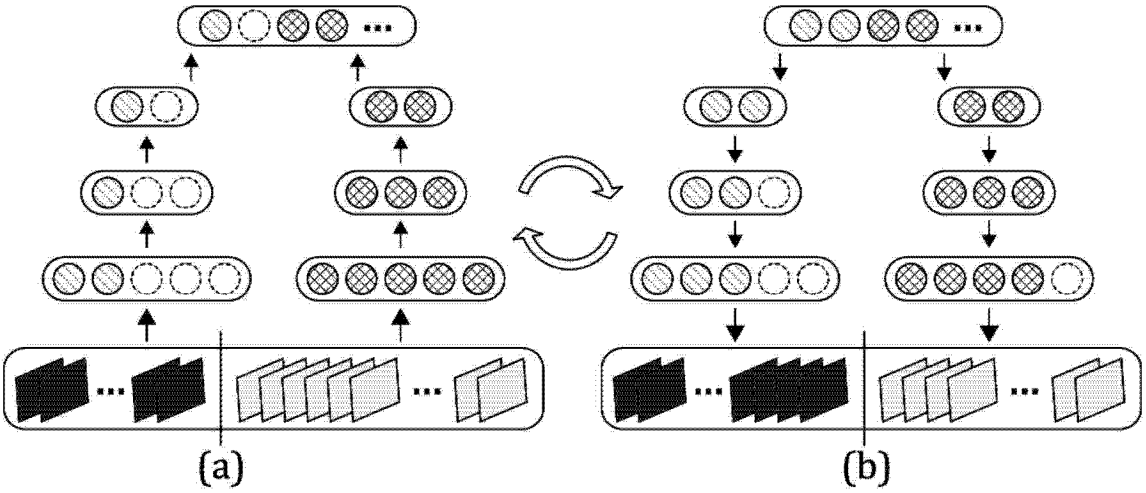


图 6