(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 110942037 A
(43)申请公布日 2020.03.31

(21)申请号 201911200563.4

(22)申请日 2019.11.29

(71)申请人 河海大学
地址 210098 江苏省南京市鼓楼区西康路1号

(72)发明人 王敏 吴敏

(74)专利代理机构 南京苏高专利商标事务所（普通合伙）32204
代理人 王安琪

(51)Int.Cl.
G06K 9/00(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

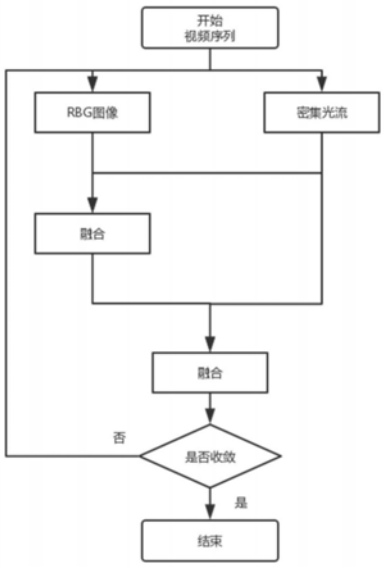
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种用于视频分析中的动作识别方法

(57)摘要

本发明公开了一种用于视频分析中的动作识别方法，包括如下步骤：(1)获取动作视频，将其处理成静止视频帧，计算叠加光流图；(2)采用步骤(1)中获得的静止图像帧数据以及光流图作为输入分别进行训练，学习特征；(3)对于步骤(2)中卷积层的时空特征进行卷积计算进行融合，并且进行3D池化，同时光流网络不截断，进行3D池化后继续提取特征；(4)将步骤(3)中得到的融合特征与光流特征进行平均计算融合；(5)根据损失函数对网络迭代训练，直至模型结果收敛。本发明能够在有限的时间规模内，尽量多的获取视频中的信息，从而增加网络的鲁棒性以及提高识别的准确率。



CN 110942037 A

1.一种用于视频分析中的动作识别方法,其特征在于,包括如下步骤:

(1)获取动作视频,将其处理成静止视频帧,计算叠加光流图;

(2)采用步骤(1)中获得的静止图像帧数据以及光流图作为输入分别进行训练,学习特征;

(3)对于步骤(2)中卷积层的时空特征进行卷积计算进行融合,并且进行3D池化,同时光流网络不截断,进行3D池化后继续提取特征;

(4)将步骤(3)中得到的融合特征与光流特征进行平均计算融合;

(5)根据损失函数对网络迭代训练,直至模型结果收敛。

2.如权利要求1所述的用于视频分析中的动作识别方法,其特征在于,步骤(1)中,计算叠加光流图具体包括如下步骤:

(11)首先计算光流图的光流矢量;对于连续帧t和t+1,它们之间的一组位移矢量场表示为d_t,在第t帧的像素点(u,v)处的位移矢量使用$d_t(u,v)$表示,它表示该像素点从第t帧移动到第t+1帧的对应像素点的位移矢量;

(12)将长度为L的连续帧矢量场的水平分量d_t^x和垂直分量d_t^y叠加起来,形成总长为2L的输入光流矢量$I_\tau \in \mathbb{R}^{H\times W\times 2L}$,用来表示连续帧间的运动信息:

$$I_\tau(u,v,2k-1)=d_{t+k-1}^x(u,v),$$

$$I_\tau(u,v,2k)=d_{t+k-1}^y(u,v),$$

其中,u=[1,W],v=[1,H],k=[1,L],W和H为视频的宽度和高度,对于任意像素点(u,v),叠加光流矢量表示为$I_\tau(u,v,c)$,c=[1,2L]是对长度为L的帧序列中该像素点运动信息的编码。

3.如权利要求1所述的用于视频分析中的动作识别方法,其特征在于,步骤(3)中,对于步骤(2)中卷积层的时空特征进行卷积计算进行融合,并且进行3D池化,同时光流网络不截断,进行3D池化后继续提取特征具体包括如下步骤:

在时间t融合两个网络中的特征图$x_t^a \in \mathbb{R}^{H\times W\times D}$和$x_t^b \in \mathbb{R}^{H'\times W'\times D'}$,其融合函数表示为$f: x_t^a,\ x_t^b \rightarrow y_t$,从而产生一张输出特征图为$y_t \in \mathbb{R}^{H''\times W''\times D''}$,其中W,H,D分别表示的是特征图的宽,高以及相对应的特征图的通道数;

(31)首先在通道d上的相同空间位置i,j堆叠两个特征图:

$y_{i,j,2d}^{cat}=x_{i,j,d}^a,\ \ y_{i,j,2d-1}^{cat}=x_{i,j,d}^b$,其中$y \in \mathbb{R}^{H\times W\times 2D}$;

(32)对于步骤(31)中得到的堆叠后的特征图$y^{cat}=f^{cat}(x^a,x^b)$,将其与过滤器$f \in \mathbb{R}^{1\times 1\times 2D\times D}$以及偏移参数$b \in \mathbb{R}^D$,在相同的空间位置i,j和特征通道d,进行卷积计算$y^{conv}=f^{conv}(x^a,x^b)$,卷积融合表示为:

$y^{conv}=y^{cat}*f+b$,

其输出结果的通道数为D,过滤器的维度是1×1×2D,其中1≤i≤H,1≤j≤W,1≤d≤D,同时,$x^a,x^b,y \in \mathbb{R}^{H\times W\times D}$,这里的过滤器f用于将维度减少2倍,并且能够在相同的空间位置上对两个特征图x^a,x^b进行加权组合;

(33)对于步骤(31)中得到的融合后的时空特征图进行3D池化,将时间t=1...T上的时

空特征图叠加起来,得到输入$x \in \mathbb{R}^{H \times W \times T \times D}$,采用大小为W′×H′×T′的池化窗口对其进行最大池化操作;

(34)对于卷积融合前的光流特征进行3D池化,同步骤(33),将2D池化扩展到时间域。

一种用于视频分析中的动作识别方法

技术领域

[0001] 本发明涉及视频分析技术领域，尤其是一种用于视频分析中的动作识别方法。

背景技术

[0002] 随着多媒体时代的到来，共享视频变得更加普遍，网络上视频的传播与获取变得越来越便捷，使得视频数据的数量急剧上升。针对数量巨大的视频数据，如何分析和利用这些数据的内容成为计算机视觉领域内的一个具有重要意义和研究价值的难题。视频分析人体动作的目标是获取视频中的图像序列，训练学习并且分析理解其中人的行为动作的含义。因此动作识别在信息获取、视频监控、人机交互等各个领域有着广泛的应用价值。

[0003] 由于卷积神经网络在计算机视觉领域内的图像应用上得到了很好的成果，由此，研究学者们将其应用于视频分析来进行动作识别中的特征提取。单是获取视频图像中的空间、纹理、背景等静态信息对于复杂的识别任务是不够的，所以需要捕捉更多的动态信息，光流能够对视频中的时间信息有效地提取，被广泛地应用于视频分析任务中。

[0004] 视频分析动作识别的重要研究内容之一是如何充分利用视频中的图像信息以及运动信息，同时这也是研究过程中亟需解决的难题。动作识别的主要目标是通过学习视频图像中人物的运动模式，将其与动作类别之间建立对应关系，从而实现理解人物的动作。因此首先需要解决如何充分提取融合视频中的图像和运动特征信息这一难点，才能够以此为基础进行后续的学习训练和分类识别。由此本发明增加了卷积融合层将提取到的时空特征融合并进行3D池化，同时不截断时间流，将训练后得到的融合时空流以及时间流再次融合，从像素水平对空间信息和时间信息建立起对应关系，从而实现更有效的特征融合。

发明内容

[0005] 本发明所要解决的技术问题在于，提供一种用于视频分析中的动作识别方法，能够在有限的时间规模内，尽量多的获取视频中的信息，从而增加网络的鲁棒性以及提高识别的准确率。

[0006] 为解决上述技术问题，本发明提供一种用于视频分析中的动作识别方法，包括如下步骤：

[0007] (1)获取动作视频，将其处理成静止视频帧，计算叠加光流图；

[0008] (2)采用步骤(1)中获得的静止图像帧数据以及光流图作为输入分别进行训练，学习特征；

[0009] (3)对于步骤(2)中卷积层的时空特征进行卷积计算进行融合，并且进行3D池化。同时光流网络不截断，进行3D池化后继续提取特征；

[0010] (4)将步骤(3)中得到的融合特征与光流特征进行平均计算融合；

[0011] (5)根据损失函数对网络迭代训练，直至模型结果收敛。

[0012] 优选的，步骤(1)中，计算叠加光流图具体包括如下步骤：

[0013] (11)首先计算光流图的光流矢量；对于连续帧t和t+1，它们之间的一组位移矢量

场表示为d_t,在第t帧的像素点(u,v)处的位移矢量使用$d_t(u,v)$表示,它表示该像素点从第t帧移动到第t+1帧的对应像素点的位移矢量;

[0014] (12)将长度为L的连续帧矢量场的水平分量d_t^x和垂直分量d_t^y叠加起来,形成总长为2L的输入光流矢量$I_\tau \in \mathbb{R}^{H\times W\times 2L}$,用来表示连续帧间的运动信息:

[0015] $$I_\tau(u,v,2k-1)=d_{t+k-1}^x(u,v),$$

[0016] $$I_\tau(u,v,2k)=d_{t+k-1}^y(u,v),$$

[0017] 其中,u=[1,W],v=[1,H],k=[1,L],W和H为视频的宽度和高度,对于任意像素点(u,v),叠加光流矢量表示为$I_\tau(u,v,c)$,c=[1,2L]是对长度为L的帧序列中该像素点运动信息的编码。

[0018] 优选的,步骤(3)中,对于步骤(2)中卷积层的时空特征进行卷积计算进行融合,并且进行3D池化,同时光流网络不截断,进行3D池化后继续提取特征具体包括如下步骤:

[0019] 在时间t融合两个网络中的特征图$x_t^a \in \mathbb{R}^{H\times W\times D}$和$x_t^b \in \mathbb{R}^{H'\times W'\times D'}$,其融合函数表示为$f: x_t^a,\ x_t^b \rightarrow y_t$,从而产生一张输出特征图为$y_t \in \mathbb{R}^{H''\times W''\times D''}$,其中W,H,D分别表示的是特征图的宽,高以及相对应的特征图的通道数;

[0020] (31)首先在通道d上的相同空间位置i,j堆叠两个特征图:

[0021] $$y_{i,j,2d}^{cat}=x_{i,j,d}^a,\quad y_{i,j,2d-1}^{cat}=x_{i,j,d}^b,$$其中$y \in \mathbb{R}^{H\times W\times 2D}$;

[0022] (32)对于步骤(31)中得到的堆叠后的特征图$y^{cat}=f^{cat}(x^a,x^b)$,将其与过滤器$f \in \mathbb{R}^{1\times 1\times 2D\times D}$以及偏移参数$b \in \mathbb{R}^D$,在相同的空间位置i,j和特征通道d,进行卷积计算$y^{conv}=f^{conv}(x^a,x^b)$,卷积融合表示为:

[0023] $$y^{conv}=y^{cat}*f+b,$$

[0024] 其输出结果的通道数为D,过滤器的维度是1×1×2D,其中1≤i≤H,1≤j≤W,1≤d≤D,同时,$x^a,x^b,y \in \mathbb{R}^{H\times W\times D}$,这里的过滤器f用于将维度减少2倍,并且能够在相同的空间位置上对两个特征图x^a,x^b进行加权组合;

[0025] (33)对于步骤(31)中得到的融合后的时空特征图进行3D池化,将时间t=1...T上的时空特征图叠加起来,得到输入$x \in \mathbb{R}^{H\times W\times T\times D}$,采用大小为W′×H′×T′的池化窗口对其进行最大池化操作;

[0026] (34)对于卷积融合前的光流特征进行3D池化,同步骤(33),将2D池化扩展到时间域。

[0027] 本发明的有益效果为:本发明将视频中的图像特征与运动特征结合起来用于识别,采用光流图提取的运动信息对于视频图像的RGB通道的缩放、更改有着不变性,能够更好地提取视频中运动物体的边缘以及中间区域的运动信息,避免网络仅被图像信息主导;卷积融合加3D池化的方式能够根据空间和时间特征的对应关系,从像素级别融合时空信息,在有限的时间规模内,尽量多地获取视频中的信息,从而增加网络的鲁棒性以及提高识别的准确率。

附图说明

[0028] 图1为本发明的方法流程示意图。

[0029] 图2为本发明的网络结构示意图。

具体实施方式

[0030] 如图1所示，一种用于视频分析中的动作识别方法，包括如下步骤：

[0031] 步骤1：获取动作视频，将其处理成静止视频帧，计算叠加光流图。计算叠加光流图具体包括以下步骤：

[0032] 步骤101：首先计算光流图的光流矢量。对于连续帧t和t+1，它们之间的一组位移矢量场表示为d_t。在第t帧的像素点(u,v)处的位移矢量使用$d_t(v,v)$表示，它表示该像素点从第t帧移动到第t+1帧的对应像素点的位移矢量。

[0033] 步骤102：将长度为L的连续帧矢量场的水平分量d_t^x和垂直分量d_t^y叠加起来，形成总长为2L的输入光流矢量$I_\tau \in \mathbb{R}^{H\times W\times 2L}$，用来表示连续帧间的运动信息：

[0034] $$I_\tau(u,v,2k-1)=d^x_{t+k-1}(u,v),$$

[0035] $$I_\tau(u,v,2k)=d^y_{t+k-1}(u,v),$$

[0036] 其中，u＝[1,W]，v＝[1,H]，k＝[1,L]，W和H为视频的宽度和高度。对于任意像素点(u,v)，叠加光流矢量$I_\tau(u,v,c)$，c＝[1,2L]是对长度为L的帧序列中该像素点运动信息的编码。

[0037] 步骤2：采用步骤1中获得的静止图像帧数据以及光流图作为输入分别提取特征。特征提取包括三层卷积和池化交替，紧接着三层卷积层以及ReLu激活函数。

[0038] 步骤3：对于步骤2中卷积层的时空特征进行卷积计算进行融合。同时光流网络不截断，进行3D池化后继续提取特征。具体包括以下步骤：

[0039] 在时间t融合两个网络中的特征图$x_t^a \in \mathbb{R}^{H\times W\times D}$和$x_t^b \in \mathbb{R}^{H'\times W'\times D'}$，其融合函数表示为f：$x_t^a,\ x_t^b \to y_t$，从而产生一张输出特征图为$y_t \in \mathbb{R}^{H''\times W''\times D''}$，其中W，H，D分别表示的是特征图的宽，高以及相对应的特征图的通道数。

[0040] 步骤301：首先在通道d上的相同空间位置i，j堆叠两个特征图：

[0041] $y^{cat}_{i,j,2d-1}=x^b_{i,j,d},\ y^{cat}_{i,j,2d}=x^a_{i,j,d}$，其中$y \in \mathbb{R}^{H\times W\times 2D}$。

[0042] 步骤302：对于步骤201中得到的堆叠后的特征图$y^{cat}=f^{cat}(x^a,x^b)$，将其与过滤器$f \in \mathbb{R}^{1\times 1\times 2D\times D}$以及偏移参数$b \in \mathbb{R}^D$，在相同的空间位置i，j和特征通道d，进行卷积计算$y^{conv}=f^{conv}(x^a,x^b)$，卷积融合表示为$y^{conv}=y^{cat}*f+6$，输出结果的通道数为D，过滤器的维度是1×1×2D，其中1≤i≤H，1≤j≤W，1≤d≤D，同时，$x^a,x^b,y \in \mathbb{R}^{H\times W\times D}$，这里的过滤器f用于将维度减少2倍，并且能够在相同的空间位置上对两个特征图x^a，x^b进行加权组合。

[0043] 步骤303：对于步骤301中得到的融合后的时空特征图进行3D池化。将时间t＝1...T上的时空特征图叠加起来，得到输入$x \in \mathbb{R}^{H\times W\times T\times D}$，采用大小为W′×H′×T′的池化窗口对其进行最大池化操作。

[0044] 步骤304:对于卷积融合前的光流特征进行3D池化,同步骤303,将2D池化扩展到时间域。

[0045] 步骤4:将步骤3中得到的融合特征与光流特征进行平均计算融合。

[0046] 步骤5:根据损失函数对网络迭代训练,直至模型结果收敛。

[0047] 损失函数采用交叉熵损失函数:$E(a,y)=-\sum_j a_j \log y_j$,其中$a_j$表示目标标签值,$y_j$表示输出值。

[0048] 网络融合能够更加充分地利用时空信息,简单的平均、相加或是最大融合对时序信息并不敏感,这说明融合过程中没有获得很多的时序信息。而卷积融合加3D池化的方式能够根据空间和时间特征的对应关系,从像素级别融合时空信息,在有限的时间规模内,尽量多地获取视频中的信息,避免网络仅被图像信息主导,从而增加网络的鲁棒性以及提高识别的准确率。同时不因为时空特征的融合而截断光流网络的特征,3D池化并用于最后的分类融合,能够将2D扩展到3D池,在长时间间隔内捕捉到同一物体的特征,进一步提高识别的准确率。

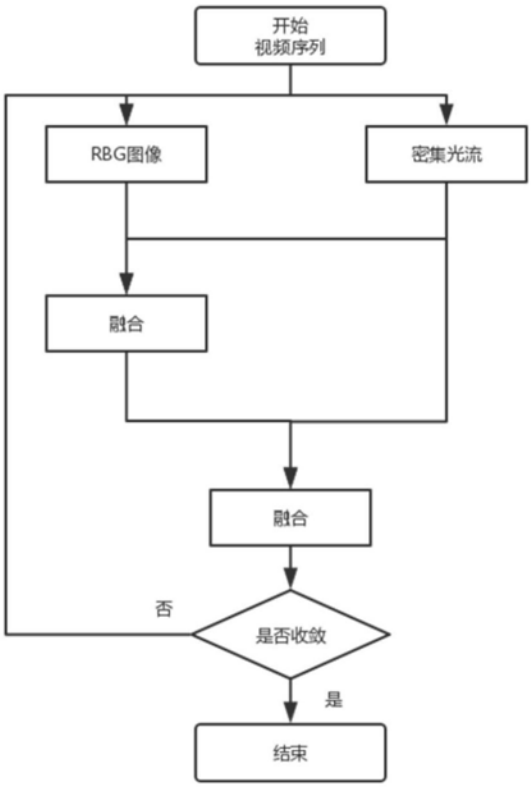


图1

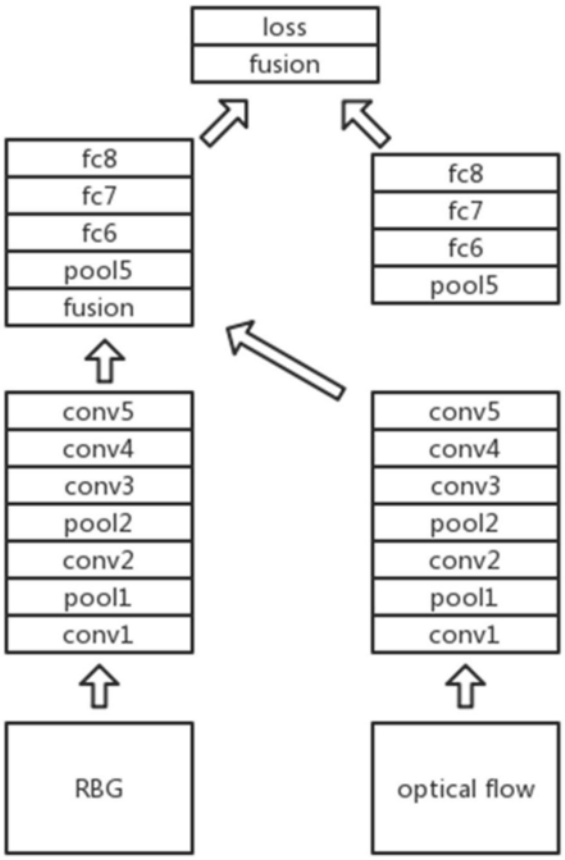

loss
fusion
fc8
fc7
fc6
pool5
fusion
fc8
fc7
fc6
pool5
conv5
conv4
conv3
pool2
conv2
pool1
conv1
conv5
conv4
conv3
pool2
conv2
pool1
conv1
RBG
optical flow

图2