(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107437083 B
(45)授权公告日 2020.09.22

(21)申请号 201710703259.6

(22)申请日 2017.08.16

(65)同一申请的已公布的文献号
申请公布号 CN 107437083 A

(43)申请公布日 2017.12.05

(73)专利权人 广西荷福智能科技有限公司
地址 530000 广西壮族自治区南宁市高新区高科路8号电子产业园1#楼厂房七层北面C座D1-35号

(72)发明人 王嘉欣 刘祎楠 王兵

(74)专利代理机构 成都华风专利事务所(普通合伙) 51223
代理人 徐丰 张巨箭

(51)Int.Cl.
G06K 9/00(2006.01)

(56)对比文件
CN 106709461 A,2017.05.24
CN 106462744 A,2017.02.22
CN 106844573 A,2017.06.13
CN 102763407 A,2012.10.31
US 2016267669 A1,2016.09.15
US 9576214 B1,2017.02.21
WO 2012167616 A1,2012.12.13
US 9152860 B2,2015.10.06
Y. Han, etal.《Video Action Recognition Based on Deeper Convolution Networks with Pair-Wise Frame Motion Concatenation》.《2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)》.2017,第1226-1235页.

审查员 祝亚尊

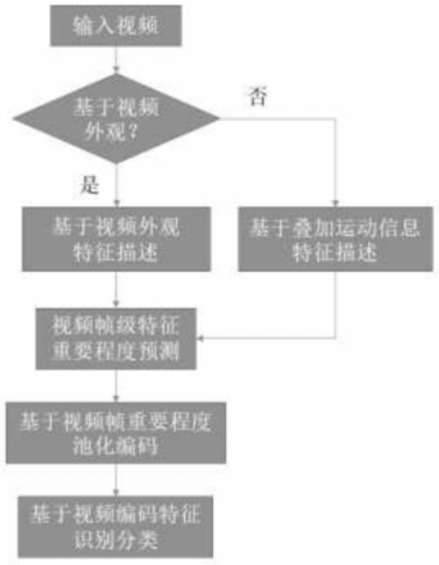
权利要求书2页 说明书4页 附图3页

(54)发明名称

一种自适应池化的视频行为识别方法

(57)摘要

本发明提供了一种自适应池化的视频行为识别方法,使用自适应池化的编码形式,可以有效利用视频中的重要信息,并将视频作为一个整体去进行相应的处理,可以对测试视频每一帧进行重要程度预测,并基于重要程度的高低,对视频帧进行池化编码操作。由于是基于当前视频中重要帧的分布来决定视频的特征编码,因此具有自适应的特性,可以对任意视频样本寻找出最适合的描述方案。



CN 107437083 B

1.一种自适应池化的视频行为识别方法,其特征在于,通过如下步骤实现:

步骤1:进行视频特征提取网络的预训练,包括基于视频外观的特征描述和基于叠加运动信息的视频特征描述,完成对于视频帧图像的特征描述;

步骤2:利用多层感知机的形式构建层数为3层的视频帧重要程度预测模块,其中运用tanh作为非线性单元,在最后一层使用sigmoid激活函数,采用基于历史累积的视频帧重要程度预测方法判断每一帧视频的重要程度,并根据其重要程度不同对视频进行池化编码,计算出视频帧的重要程度以及积累到某一时刻的视频联合特征描述;

步骤3:将得到的视频联合特征描述,进行二范数归一化,并使用全连接层对视频样本进行类别预测,采用标准交叉熵计算预测类别与真实类别之间的损失。

2.根据权利要求1所述的自适应池化的视频行为识别方法,其特征在于:所述基于视频外观的特征描述的具体操作方法为,将视频帧图像尺寸归一化为224*224*3,并输入到在ImageNet预训练的VGG-16网络结构,提取第一层全连接层(FC-6)的输出作为每一帧视频的特征描述。

3.根据权利要求1所述的自适应池化的视频行为识别方法,其特征在于:所述基于叠加运动信息的视频特征描述的具体操作方法为,首先计算出视频中任意相邻两帧之间的运动信息,并将连续10帧的水平方向运动以及垂直方向运动相叠加构成运动谱,之后将叠加运动信息尺寸归一化为224*224*20,然后将其输入到在UCF101视频数据库进行预训练的VGG-16网络结构,提取第一层全连接层(FC-6)的输出作为对应视频帧的特征描述,采用该方法时,需要对连续20帧视频图像进行叠加,从视频末尾19帧开始,就不在构建此类特征表示。

4.根据权利要求1所述的自适应池化的视频行为识别方法,其特征在于:所述基于历史累积的视频帧重要程度预测方法通过如下方式进行计算:

基于第1帧至第t帧,通过公式$\rho_{t+1} = f_{imp}(\psi(X,t),\varphi(\boldsymbol{x}_{t+1}))$得到视频第t+1帧的重要程度预测,其中$f_{imp}$为多层感知机结构,层数为3层,$\varphi(\boldsymbol{x}_{t+1})$是第t+1帧的视频帧图像特征;

当得到第t+1帧的视频帧重要程度之后,利用平均池化计算方法计算出从第1帧至第t+1帧的联合特征描述形式,计算公式由下式给出:

$$\psi(X,t+1) = (1/\widehat{\rho}_{t+1})(\widehat{\rho}_t\psi(X,t) + \rho_{t+1}\varphi(\boldsymbol{x}_{t+1}))。$$

其中,山(X,t+1)表示为视频中第1帧至第t+1帧的联合特征描述,$\widehat{\rho}_{t+1}$表示视频中第1帧至第t+1帧的重要程度之和,$\widehat{\rho}_t$表示视频中第1帧至第t帧的重要程度之和,$\widehat{\rho}_t$的计算方式为$\widehat{\rho}_t = \sum_{k=1}^{t}\rho_k$。

5.根据权利要求1所述的自适应池化的视频行为识别方法,其特征在于:在训练阶段,预测类别与真实类别之间的损失,由损失函数公式:

$L(X,y) = L_{CE}(X,y) + \lambda L_E(\tau)$

给出,其中,$L_{CE}(X,y)$表示视频样本预测类别与真实类别之间的交叉熵损失,λ表示约束项$L_E(\tau)$的调节参数,当λ选择过大时,只有较少的视频帧参与分类,增大分类难度;当λ选择过小时,会有较多视频帧参与分类,此时对训练数据使用多尺度随机缩放移位翻转的操作来避免过拟合,$L_E(\tau)$用来保证选择到的视频帧都是重要程度较大的视频帧,计算方法下式

给出：

$$L_E(\tau) = -\sum_k (e^{\rho_k}/N)\log(e^{\rho_k}/N)$$

其中，N由公式$N = \sum_t e^{\rho_t}$计算得出，ρ_t表示第t帧的重要程度。

6.根据权利要求1所述的自适应池化的视频行为识别方法，其特征在于：训练时初始学习率设置为0.001并在每3000次迭代后递减为上一组的10％，优化算法选择为随机梯度下降算法。

7.根据权利要求1所述的自适应池化的视频行为识别方法，其特征在于：分类网络的初始学习率设置为10^{-6}，使用adam算子进行优化。

一种自适应池化的视频行为识别方法

技术领域

[0001] 本发明提出一种自适应池化的视频行为识别方法，是一种用于对视频样本进行分类识别的新型应用技术。

背景技术

[0002] 随着智能手机以及互联网的普及，视频分享的数量得到了指数级的增长。其中大量网络视频的被拍摄目标都是人的某种行为活动。视频行为识别主要目的是将海量视频文件进行有效的分类，以方便人们可以快速提取出自己感兴趣的视频文件。视频行为识别已成为一个热门的领域，具有广泛的应用场景，如：异常目标行为分类、视频内容理解、家庭监控视频检测、视频推荐等。

[0003] 在计算机视觉领域，有诸多方法可以用于视频行为识别。这些方法主要可以分为两类：一类是基于传统数据驱动特征设计的视频行为识别方法，另一类是基于神经网络深度学习的视频行为识别方法。

[0004] 传统的数据驱动特征设计方法，主要基于视频自身数据的内容，结合开发者人工设计的数据描述形式。比如：利用视频同一帧内像素之间关系构建梯度方向直方图(HoG)、利用视频不同帧间运动信息构建的光流方向直方图(HoF)、利用运动信息的一阶导数构建的运动边界直方图(MBH)等等描述作为低层特征，并使用词袋模型(BoW)或费舍尔向量去进行视频级的特征描述，最终使用支持向量机(SVM)、随机森林(RF)等分类工具对视频样本进行训练以及分类操作。

[0005] 相比传统数据驱动特征设计方法，基于神经网络深度学习的方法，如双路深度神经网络(two-streamDCNN)可以利用大量视频训练样本训练得到一系列针对视频内容本身的特征滤波器。常见的如：使用视频帧所训练得到的空域网络(spatial-streamDCNN)以及使用视频多叠加运动信息训练得到的时域网络(temporal-streamDCNN)。可以从视频的外观信息以及运动信息两个层面去更好的提取视频内在特征，并产生更好的分类效果。

[0006] 基于神经网络深度学习的方法尽管有着传统数据驱动设计所不具备的优势。但是在对视频样本的描述时也会存在一些问题。由于邻近的视频帧图像内容比较相似，所以会产生大量冗余信息，一般基于神经网络深度学习的方法无法区分任意视频帧对于当前视频分类的重要程度。与此同时，此类方法在分类视频时，本质上是在完成多次图像分类的操作，并没有将视频当成一个整体去处理。

[0007] 为了应对上述提出的两个主要问题，本文发明提出了一种自适应池化的视频行为识别方法。该方法可以有效对视频中的每一帧图像给出重要程度的判断，并且可以将判断为比较重要的视频帧使用池化的方式编码为视频样本的特征描述并进行分类。相比于传统的神经网络深度学习方案，本方法可以更为有效的发现视频中哪些信息才是重要信息，并对这些信息加以更充分的利用。

发明内容

[0008] 为解决上述技术问题，本发明采用的一个技术方案是：一种自适应池化的视频行为识别方法，其特征在于，通过如下步骤实现：

[0009] 步骤1：进行视频特征提取网络的预训练，包括基于视频外观的特征描述和基于叠加运动信息的视频特征描述，完成对于视频帧图像的特征描述；

[0010] 步骤2：利用多层感知机的形式构建层数为3层的视频帧重要程度预测模块，其中运用tanh作为非线性单元，在最后一层使用sigmoid激活函数，采用基于历史累积的视频帧重要程度预测方法判断每一帧视频的重要程度，并根据其重要程度不同对视频进行池化编码，计算出视频帧的重要程度以及积累到某一时刻的视频联合特征描述；

[0011] 步骤3：将得到的视频联合特征描述，进行二范数归一化，并使用全连阶层对视频样本进行类别预测，采用标准交叉熵计算预测类别与真实类别之间的损失。

[0012] 具体而言，对于视频帧图像特征描述，我们采用两种方法：(1)基于视频外观信息进行描述，此时我们将视频帧图像尺寸归一化为224*224*3，并输入到在ImageNet预训练的VGG-16网络结构，提取第一层全连接层(FC-6)的输出作为每一帧视频的特征描述。(2)基于视频运动信息进行描述，我们首先计算出视频中任意相邻两帧之间的运动信息，并将连续10帧的水平方向运动以及垂直方向运动相叠加构成运动谱，之后将叠加运动信息尺寸归一化为224*224*20，然后将其输入到在UCF101视频数据库进行预训练的VGG-16网络结构，提取第一层全连接层(FC-6)的输出作为对应视频帧的特征描述。

[0013] 对于自适应池化编码，考虑到视频本质上是一系列连续运动的图像。本发明设计了一种基于历史积累的视频帧重要程度预测方法。

[0014] 该方法中，我们对视频第t+1帧的预测是基于第1帧至第t帧共同得到。对第t+1帧重要程度的预测由公式(1)给出：

[0015] $$\rho_{t+1} = f_{imp}(\psi(X,t), \varphi(\boldsymbol{x}_{t+1})) \quad (1)$$

[0016] 其中，f_{imp}为多层感知机结构，层数为3.$\varphi(\boldsymbol{x}_{t+1})$是第t+1帧的视频帧图像特征。

[0017] 当我们得到第t+1帧的视频帧重要程度之后，我们可以计算出从第1帧至第t+1帧的联合特征描述形式，平均池化计算方法由公式(2)给出：

[0018] $$\psi(X,t+1) = (1/\hat{\rho}_{t+1})(\hat{\rho}_t \psi(X,t) + \rho_{t+1}\varphi(\boldsymbol{x}_{t+1})) \quad (2)$$

[0019] 其中，ψ(X,t+1)表示为视频中第1帧至第t+1帧的联合特征描述，$\hat{\rho}_{t+1}$表示视频中第1帧至第t+1帧的重要程度之和，$\hat{\rho}_t$表示视频中第1帧至第t帧的重要程度之和，计算方法由公式(3)给出：

[0020] $$\hat{\rho}_t = \sum_{k=1}^{t} \rho_k \quad (3)$$

[0021] 公式(3)中ρ_k表示第k帧的重要程度预测。

[0022] 对于视频样本类别预测，本发明使用标准交叉熵去计算预测类别与真实类别之间的损失，在训练阶段，损失函数由公式(4)给出：

[0023] $L(X,y) = L_{CE}(X,y) + \lambda L_E(\tau)$ (4)

[0024] 其中，$L_{CE}(X,y)$表示视频样本预测类别与真实类别之间的交叉熵损失，λ表示约束项$L_E(\tau)$的调节参数，当λ选择过大时，只有较少的视频帧参与分类，增大分类难度；当λ选择

过小时，会有较多视频帧参与分类，但此时容易使训练过拟合，根据具体实验效果，我们将λ设置为10^5.$L_E(\tau)$用来保证选择到的视频帧都是重要程度较大的视频帧，计算方法由公式(5)给出：

[0025] $$L_E(\tau) = -\sum_k (e^{\rho_k}/N)\log((e^{\rho_k}/N)) \quad (5)$$

[0026] 其中N的计算方法由公式(6)给出：

[0027] $N = \sum_t e^{\rho t}$ (6)

[0028] 其中ρ_t表示第t帧的重要程度。

[0029] 区别于现有技术的情况，本发明的有益效果是：

[0030] 首先本发明可以应对多种视频帧级特征作为输入，有着广泛的应用环境。其次，本方法可以有效的对视频每一帧进行重要程度预测，既可以有效筛选视频中的重要信息用以描述视频，也可以用以发现视频中的感兴趣信息。同时，利用对视频所有帧重要程度的预测，本文使用池化编码的形式，对不同视频实现了自适应的视频描述方案。并且最终将视频重要程度预测、池化编码以及视频识别分类模块内嵌到一个端到端的网络结构，从而实现了可以快速进行训练测试以及调试的工作。本发明提出的方法在公开数据集上进行测试，验证了方法的有效性。

附图说明

[0031] 图1是本发明实施流程图。

[0032] 图2是本发明自适应池化编码模块示意图。

[0033] 图3是本发明基于视频外观信息特征提取示意图。

[0034] 图4是本发明基于视频运动信息特征提取示意图。

具体实施方式

[0035] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅是本发明的一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0036] 参见图1提供的一种自适应池化的视频行为识别方法，通过如下步骤实现：

[0037] 步骤1：进行视频特征提取网络的预训练，包括基于视频外观的特征描述和基于叠加运动信息的视频特征描述，完成对于视频帧图像的特征描述；

[0038] 步骤2：利用多层感知机的形式构建层数为3层的视频帧重要程度预测模块，其中运用tanh作为非线性单元，在最后一层使用sigmoid激活函数，采用基于历史累积的视频帧重要程度预测方法判断每一帧视频的重要程度，并根据其重要程度不同对视频进行池化编码，计算出视频帧的重要程度以及积累到某一时刻的视频联合特征描述；

[0039] 步骤3：将得到的视频联合特征描述，进行二范数归一化，并使用全连阶层对视频样本进行类别预测，采用标准交叉熵计算预测类别与真实类别之间的损失。

[0040] 具体而言，

[0041] 如图3、图4所示，对于视频帧图像特征描述，我们采用两种方法：(1)基于视频外观信息进行描述，此时我们将视频帧图像尺寸归一化为224*224*3，并输入到在ImageNet预训

练的VGG-16网络结构，提取第一层全连接层(FC-6)的输出作为每一帧视频的特征描述。(2)基于视频运动信息进行描述，我们首先计算出视频中任意相邻两帧之间的运动信息，并将连续10帧的水平方向运动以及垂直方向运动相叠加构成运动谱，之后将叠加运动信息尺寸归一化为224*224*20，然后将其输入到在UCF101视频数据库进行预训练的VGG-16网络结构，提取第一层全连接层(FC-6)的输出作为对应视频帧的特征描述。

[0042] 对于自适应池化编码，如图2所示，考虑到视频本质上是一系列连续运动的图像，本发明设计了一种基于历史积累的视频帧重要程度预测方法。

[0043] 该方法中，我们对视频第t+1帧的预测是基于第1帧至第t帧共同得到。对第t+1帧重要程度的预测由公式(1)给出：

[0044] $$\rho_{t+1} = f_{imp}(\psi(X,t), \varphi(\boldsymbol{x}_{t+1})) \quad (1)$$

[0045] 其中，f_{imp}为多层感知机结构，层数为3.$\varphi(\boldsymbol{x}_{t+1})$是第t+1帧的视频帧图像特征。

[0046] 当我们得到第t+1帧的视频帧重要程度之后，我们可以计算出从第1帧至第t+1帧的联合特征描述形式，平均池化计算方法由公式(2)给出：

[0047] $$\psi(X,t+1) = (1/\hat{\rho}_{t+1})(\hat{\rho}_t \psi(X,t) + \rho_{t+1}\varphi(\boldsymbol{x}_{t+1})) \quad (2)$$

[0048] 其中，ψ(X,t+1)表示为视频中第1帧至第t+1帧的联合特征描述，$\hat{\rho}_{t+1}$表示视频中第1帧至第t+1帧的重要程度之和，$\hat{\rho}_t$表示视频中第1帧至第t帧的重要程度之和，计算方法由公式(3)给出：

[0049] $$\hat{\rho}_t = \sum_{k=1}^{t} \rho_k \quad (3)$$

[0050] 公式(3)中ρ_k表示第k帧的重要程度预测。

[0051] 对于视频样本类别预测，本发明使用标准交叉熵去计算预测类别与真实类别之间的损失，在训练阶段，损失函数由公式(4)给出：

[0052] $L(X,y) = L_{CE}(X,y) + \lambda L_E(\tau)$ (4)

[0053] 其中，$L_{CE}(X,y)$表示视频样本预测类别与真实类别之间的交叉熵损失，λ表示约束项$L_E(\tau)$的调节参数，当λ选择过大时，只有较少的视频帧参与分类，增大分类难度；当λ选择过小时，会有较多视频帧参与分类，但此时容易使训练过拟合，根据具体实验效果，我们将λ设置为10^5.$L_E(\tau)$用来保证选择到的视频帧都是重要程度较大的视频帧，计算方法由公式(5)给出：

[0054] $$L_E(\tau) = -\sum_k (e^{\rho_k}/N)\log(e^{\rho_k}/N) \quad (5)$$

[0055] 其中N的计算方法由公式(6)给出：

[0056] $N = \sum_t e^{\rho t}$ (6)

[0057] 其中ρ_t表示第t帧的重要程度。

[0058] 以上所述仅为本发明的实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

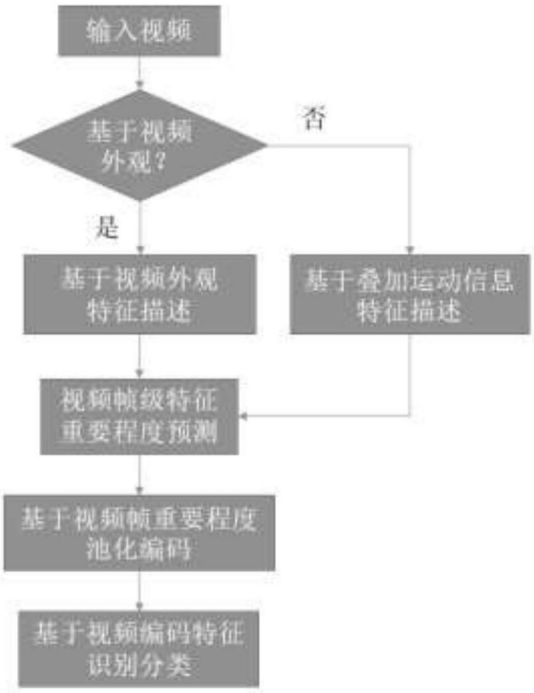

输入视频
基于视频外观？
否
是
基于视频外观特征描述
基于叠加运动信息特征描述
视频帧级特征重要程度预测
基于视频帧重要程度池化编码
基于视频编码特征识别分类

图1

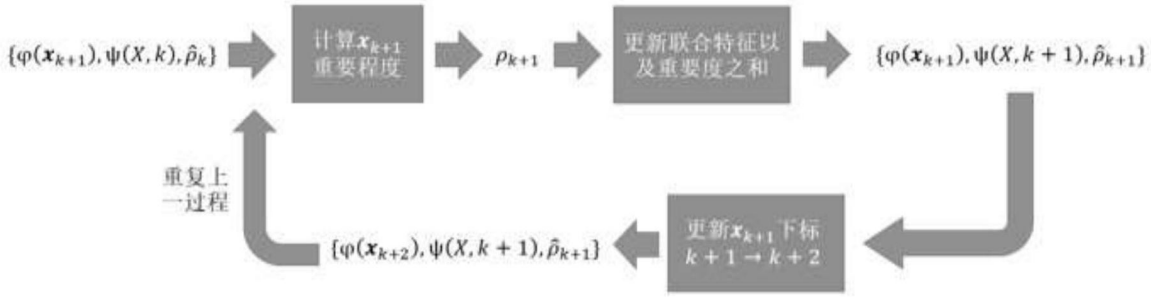

$\{\varphi(\boldsymbol{x}_{k+1}), \psi(X,k), \hat{\rho}_k\}$
计算$\boldsymbol{x}_{k+1}$重要程度
ρ_{k+1}
更新联合特征以及重要度之和
$\{\varphi(\boldsymbol{x}_{k+1}), \psi(X,k+1), \hat{\rho}_{k+1}\}$
更新$\boldsymbol{x}_{k+1}$下标 $k+1 \rightarrow k+2$
$\{\varphi(\boldsymbol{x}_{k+2}), \psi(X,k+1), \hat{\rho}_{k+1}\}$
重复上一过程

图2


224*224*RGB
Conv1_1
Conv1_2
pool1
Conv2_1
Conv2_2
pool2
Conv3_1
Conv3_2
Conv3_3
pool3
Conv4_1
Conv4_2
Conv4_3
pool4
Conv5_1
Conv5_2
Conv5_3
pool5
fc6
视频帧
特征提取模块
视频帧特征向量
$\varphi(x_p) \in R^{4096}$

图3


224*224*UV*10
Conv1_1
Conv1_2
pool1
Conv2_1
Conv2_2
pool2
Conv3_1
Conv3_2
Conv3_3
pool3
Conv4_1
Conv4_2
Conv4_3
pool4
Conv5_1
Conv5_2
Conv5_3
pool5
fc6
视频叠加运动信息
特征提取模块
视频叠加运动信息特征向量
$\varphi(x_p) \in R^{4096}$

图4