(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 112149645 A
(43)申请公布日 2020.12.29

(21)申请号 202011248793.0

(22)申请日 2020.11.10

(71)申请人 西北工业大学
地址 710072 陕西省西安市友谊西路127号

(72)发明人 王鹏 田磊

(74)专利代理机构 西北工业大学专利中心 61204
代理人 刘新琼

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/46(2006.01)
G06T 3/40(2006.01)
G06T 7/12(2017.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

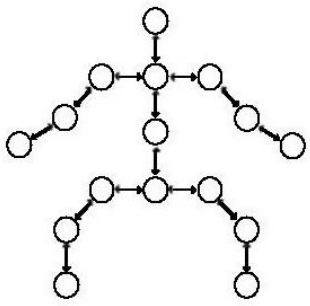
权利要求书1页 说明书6页 附图1页

(54)发明名称

基于生成对抗学习和图神经网络的人体姿势关键点识别方法

(57)摘要

本发明涉及一种生成对抗学习和图神经网络的人体姿势关键点识别方法，属于人体姿势关键点识别领域。一方面采用卷积网络作为生成器提取图片特征，然后结合反卷积学习其中的人体姿势关键点，另一方面采用图神经网络作为判别器对学习生成的人体姿势关键点进行正误判别，促使生成器加强对错误的关键点再学习，以适应更复杂环境下的人体姿势关键点识别。



CN 112149645 A

1.一种基于生成对抗学习和图神经网络的人体姿势关键点识别方法，其特征在于步骤如下：

步骤1:输入为一张含有人体姿势的图片，大小为3*256*256，表示为$V\in R^{C\times W\times H}$，其中C表示的是图像channel的大小，W和H是图像的宽和高，然后经过5层包含残差的卷积神经网络，得到256个8*8大小的特征图，接着将此256个8*8大小的特征图经过三层反卷积层放大和一层卷积神经网络得到相对应的人体姿势预测关节点的节点信息，即16个64*64大小的特征图；最后，提取每个64*64大小的特征图中的最大值作为人体关节点坐标，此处人体关节点坐标总共是16个；

步骤2:将步骤1中得到的16个64*64大小的特征图作为判别器的输入，判别器用来判断生成器生成的当前预测节点是否符合人为先验是否合理，如果合理即为1，否则为0；具体过程为:通过对输入的16个64*64大小特征图后两维进行拉伸得到16个长度为64*64的向量，经过全连接层的处理得到16个长度为256的向量，分别对每个关节点过门控图神经网络GGNN来得到更新后的节点信息，最后通过全连接层处理得到16个一维向量，即人体姿势关键点。

2.根据权利要求1所述的一种基于生成对抗学习和图神经网络的人体姿势关键点识别方法，其特征在于步骤2中所述的门控图神经网络GGNN的更新过程:第一，依靠自建的人体姿势图结构和公式(1)，得到每个节点和相邻节点构成的边邻域信息j;第二，结合每个节点(t-1)时刻的状态信息i和边邻域信息j经过公式(3)得到更新后的节点信息：

$$j_n^t = F(i_m^{t-1} \mid m \in M_n) \quad (1)$$

$$j_n^t = \sum_{n,n' \in \Omega} W_{n,n} i_n^{t-1} \quad (2)$$

$$i_n^t = GRU(i_n^{t-1},\ j_n^t) \quad (3)$$

公式(1)中n表示某个关节点，M是第n个关节点的邻域关节点集合，m表示其中某个邻域节点，t是当前更新时步，i为该节点状态信息，j为每个节点和相邻节点构成的边邻域信息，F和GRU分别表示从相邻节点收集信息和更新节点隐藏状态信息的函数，F可以表示为公式(2)，GRU可以用公式(4)-(7)表示；

门控机制GRU的具体计算公式如下：

$$z_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (4)$$

$$r_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (5)$$

$$\tilde{h}_n^t = \tanh(W_n j_n^t + U_n(i_n^{t-1} \bullet r_n^{t-1} + b) \quad (6)$$

$$i_n^t = (1 - z_n^t) \bullet i_n^{t-1} + z_n^t \bullet \tilde{h}_n^t \quad (7)$$

其中，W和U都是第n个关键点的卷积权重，b是卷积偏置；sigmoid和tanh为常用的激活函数。

基于生成对抗学习和图神经网络的人体姿势关键点识别方法

技术领域

[0001] 本发明属于人体姿势关键点识别领域，具体是提出一种结合生成对抗学习和图神经网络的人体姿势关键点识别方法和系统。整个系统一方面采用resnet卷积网络作为生成器提取图片特征，然后结合反卷积学习其中的人体姿势关键点，另一方面采用图神经网络作为判别器对学习生成的人体姿势关键点进行正误判别，促使生成器加强对错误的关键点再学习，以适应更复杂环境下的人体姿势关键点识别。

背景技术

[0002] 人体姿态关键点识别是计算机视觉领域的基本研究方向之一，在传统算法遭遇瓶颈之时，卷积神经网络的再次崛起和快速迭代为解决这一问题带来了新工具，最近几年，尽管人体姿势关键点识别任务在使用深度卷积神经网络的情况下已经取得了极大的进步，但是由于光照、遮挡和变化大的身体姿势等导致关键点不可见的问题，2D人体姿势关键点识别仍然是一项具有挑战性和重要意义的任务。人体姿势关键点识别任务广泛的应用于行为动作识别、人机交互和游戏动画等相关任务中，该任务的主要难点是识别手臂的肘和腕以及腿部的踝和膝盖。

[0003] 人体姿势识别任务中人体不同关节部分的相关空间语义信息起到非常关键的作用，考虑到人体各个关节本身部分就可以看作是一个连接的图结构，本发明采用图神经网络作为生成对抗学习中的判别器对人体各个关节部分的关键点识别的正误判别。

发明内容

[0004] 要解决的技术问题

[0005] 为了避免现有技术的不足之处，本发明提出一种基于生成对抗学习和图神经网络的人体姿势关键点识别方法。

[0006] 技术方案

[0007] 一种基于生成对抗学习和图神经网络的人体姿势关键点识别方法，其特征在于步骤如下：

[0008] 步骤1:输入为一张含有人体姿势的图片，大小为3*256*256，表示为$V \in R^{C \times W \times H}$，其中C表示的是图像channel的大小，W和H是图像的宽和高，然后经过5层包含残差的卷积神经网络，得到256个8*8大小的特征图，接着将此256个8*8大小的特征图经过三层反卷积层放大和一层卷积神经网络得到相对应的人体姿势预测关节点的节点信息，即16个64*64大小的特征图；最后，提取每个64*64大小的特征图中的最大值作为人体关节点坐标，此处人体关节点坐标总共是16个；

[0009] 步骤2:将步骤1中得到的16个64*64大小的特征图作为判别器的输入，判别器用来判断生成器生成的当前预测节点是否符合人为先验是否合理，如果合理即为1，否则为0；具体过程为:通过对输入的16个64*64大小特征图后两维进行拉伸得到16个长度为64*64的向量，经过全连接层的处理得到16个长度为256的向量，分别对每个关节点过门控图神经网络

GGNN来得到更新后的节点信息,最后通过全连接层处理得到16个一维向量,即人体姿势关键点。

[0010] 步骤2中所述的门控图神经网络GGNN的更新过程:第一,依靠自建的人体姿势图结构和公式(1),得到每个节点和相邻节点构成的边邻域信息j;第二,结合每个节点(t-1)时刻的状态信息i和边邻域信息j经过公式(3)得到更新后的节点信息:

[0011] $$j_n^t = F(i_m^{t-1} \mid m \in M_n) \quad (1)$$

[0012] $$j_n^t = \sum_{n,n' \in \Omega} W_{n,n'} i_n^{t-1} \quad (2)$$

[0013] $$i_n^t = GRU(i_n^{t-1},\ j_n^t) \quad (3)$$

[0014] 公式(1)中n表示某个关节点,M是第n个关节点的邻域关节点集合,m表示其中某个邻域节点,t是当前更新时步,i为该节点状态信息,j为每个节点和相邻节点构成的边邻域信息,F和GRU分别表示从相邻节点收集信息和更新节点隐藏状态信息的函数,F可以表示为公式(2),GRU可以用公式(4)-(7)表示;

[0015] 门控机制GRU的具体计算公式如下:

[0016] $$z_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (4)$$

[0017] $$r_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (5)$$

[0018] $$\tilde{h}_n^t = \tanh(W_n j_n^t + U_n(i_n^{t-1} \bullet r_n^{t-1} + b) \quad (6)$$

[0019] $$i_n^t = (1 - z_n^t) \bullet i_n^{t-1} + z_n^t \bullet \tilde{h}_n^t) \quad (7)$$

[0020] 其中,W和U都是第n个关键点的卷积权重,b是卷积偏置;sigmoid和tanh为常用的激活函数。

[0021] 有益效果

[0022] 本发明提出的一种基于生成对抗学习和图神经网络的人体姿势关键点识别方法,可以得到更稳定更精确的人体姿势关键点,基于图神经网络的结构充分利用了人体姿势本身内在的语义空间结构关系,结合生成对抗式的学习可以应对更多复杂的环境和变换大的姿势,而在本发明应用时不需要判别器部分,仅仅使用生成器生成所需的结果即可,如此使得网络更简单高效,运行速度更快。

附图说明

[0023] 图1图结构

[0024] 图2生成器结构图

[0025] 图3判别器结构图

具体实施方式

[0026] 现结合实施例、附图对本发明作进一步描述:

[0027] 本发明的技术方案主要分为两个模块:第一个模块是生成器(如图2),第二个模块

是判别器(如图3)。

[0028] 生成器结构:输入为3*256*256的图像,表示为$V \in R^{C \times W \times H}$,此处的C表示的是图像channel的大小,W和H是图像的宽和高,经过多层卷积神经网络得到256*8*8的特征图(feature map),此处主要是提取图片特征信息的主干网络。将此256*8*8大小的特征图经过三层反卷积层(Deconv)放大得到256*64*64的特征图,最后通过一层输出卷积得到相对应的预测关节点的节点信息,即16*64*64的特征图,此处的16为人体关节点数量。

[0029] 判别器结构:判别器的输入是生成器输出的16*64*64特征图,首先将64*64的两维特征转换为一维特征,经过一层全连接神经网络变为16*256的特征大小。通过人体关节自身的空间语义信息构建图结构(如图2),利用图结构的关系对于每个节点加上相邻节点的特征信息,得到的仍然是16*256的特征。最后通过一层全连接神经网络得到16*1的一个向量特征。

[0030] 图结构的构建:利用人体姿势本身的依赖关系构建图结构(如图1所示),具体为:将人体姿势的16个关节点作为图结构中的节点,将人体姿势的每个关节点和相邻关节点的依赖关系作为图结构中节点和节点的连接。

[0031] 端到端的训练过程:在随机初始化所有参数后,按照传统的生成对抗网络一般训练过程交替训练生成器和判别器。具体来说,生成器训练3次,判别器训练1次。在训练判别器的过程中,我们把真实的标签作为判别器的输入,让判别器来学习这是真的。同时,本发明将生成器生成的预测结果作为判别器的输入,训练判别器来学习这是假的。在训练生成器的过程中,通过生成对抗学习直接优化生成器来欺骗判别器。换句话说,判别器将把生成器产生的预测结果视为真实的结果。最后通过加权结合两部分的损失值生成对抗性的学习,用第二个模块辅助确保第一个模块有能力对各种复杂环境下的大姿势实现更稳定更精准的人体关键点定位。

[0032] 测试过程:在测试时,只需要用到生成器的输出作为最终结果即可,本身的判别器只用做训练部分来提高生成器的预测能力,测试部分不需要用到,很显然,本发明设计具有诸如速度快、模型结构简单、参数量少等多个优点。

[0033] 该人体姿势关键点识别方法有以下主要步骤:

[0034] (1)生成器的训练:将一张图片通过生成器提取特征并输出得到相对应的预测关节点的节点信息。具体过程为:输入为一张含有人体姿势的图片,大小为3*256*256,表示为$V \in R^{C \times W \times H}$,此处的C表示的是图像channel的大小,W和H是图像的宽和高,然后经过5层包含残差的卷积神经网络,得到256个8*8大小的特征图(feature map),接着将此256个8*8大小的特征图经过三层反卷积层(Deconv)放大和一层卷积神经网络得到相对应的人体姿势预测关节点的节点信息,即16个64*64大小的特征图。最后,提取每个64*64大小的特征图中的最大值作为人体关节点坐标,此处人体关节点坐标总共是16个。

[0035] (2)判别器的训练:将(1)中得到的16个64*64大小的特征图作为判别器的输入,判别器用来判断生成器生成的当前预测节点是否符合人为先验是否合理,如果合理即为1,否则为0。具体过程为:通过对输入的16个64*64大小特征图后两维进行拉伸得到16个长度为64*64的向量,经过全连接层的处理得到16个长度为256的向量,分别对每个关节点过门控图神经网络(步骤3)来得到更新后的节点信息,最后通过全连接层处理得到16个一维向量。

[0036] (3)门控图神经网络(GGNN)的更新过程:第一,依靠自建的人体姿势图结构和公式

(1),可以得到每个节点和相邻节点构成的边邻域信息j;第二,结合每个节点(t-1)时刻的状态信息i和边邻域信息j经过公式(3)得到更新后的节点信息。

[0037] $$j_n^t = F(i_m^{t-1} \mid m \in M_n) \quad (1)$$

[0038] $$j_n^t = \sum_{n,n' \in \Omega} W_{n,n'} i_n^{t-1} \quad (2)$$

[0039] $$i_n^t = GRU(i_n^{t-1},\ j_n^t) \quad (3)$$

[0040] 公式(1)中n表示某个关节点,M是第n个关节点的邻域关节点集合,m表示其中某个邻域节点,t是当前更新时步,i为该节点状态信息,j为每个节点和相邻节点构成的边邻域信息,F和GRU分别表示从相邻节点收集信息和更新节点隐藏状态信息的函数,F可以表示为公式(2),GRU可以用公式(4)-(7)表示。

[0041] (4)门控机制(GRU)的具体计算公式如下:

[0042] $$z_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (4)$$

[0043] $$r_n^t = \mathrm{sigmoid}(W_n j_n^t + U_n i_n^{t-1} + b) \quad (5)$$

[0044] $$\tilde{h}_n^t = \tanh(W_n j_n^t + U_n (i_n^{t-1} \bullet r_n^{t-1} + b) \quad (6)$$

[0045] $$i_n^t = (1 - z_n^t) \bullet i_n^{t-1} + z_n^t \bullet \tilde{h}_n^t) \quad (7)$$

[0046] 这里的W和U都是第n个关键点的卷积权重,b是卷积偏置。sigmoid和tanh为常用的激活函数。

[0047] 本发明提供了一种基于结合生成对抗学习和图神经网络的人体姿势关键点识别方法,具体过程如下:

[0048] 1、数据预处理

[0049] 给定一张包含人体姿势的图片,根据图片中人体的边界框把人裁剪出来,然后使用双线性插值的方法将图片尺寸大小调整到256×256,在裁剪和调整图片的同时需要对相应的关键点真实标签做处理。

[0050] 2、数据增强

[0051] 将同一张图片随机尺度缩放、随机左右翻转以及随机旋转一定角度θ∈[-30°,+30°],使用双线性插值的方法调整图片大小到256×256,最后归一化处理得到张量256×256×3。在图片处理变为张量256×256×3作为输入后,而图像上对应人体姿势关键点的坐标也要做相应变化。图像在左右翻转时,人体姿势左边点的坐标需要和对应的右边点的坐标交换,随机尺度缩放、随机旋转和图像大小调整时的关键点坐标也要做相应变换调整。

[0052] 3、生成器网络模块训练

[0053] 输入图片经过数据预处理后变为(256×256×3)张量,然后输入张量到Resnet网络,去掉Resnet网络最后的两层即平均池化层和全连接层,在网络后面增加三个反卷积层和一层卷积层,得到网络输出的特征图,此时的特征图大小为64×64。输出的特征图的个数即是人体姿势关键点的数量,关键点数设置为16,即输出16个关键点的坐标,然后根据这16个关键点的坐标来编码生成64×64的热图(heatmap),然后与真实标签对应的热力图(64×64)计算归一化平均误差。训练时使用Adam优化器来更新参数。

[0054] 4、图网络构建

[0055] 根据人体姿势构建图结构，如图2所示。图神经网络需要图(图被表示为G＝{I,E})作为它的输入，其中I和E分别表示为图的节点和边，每个节点i∈I拥有自身的隐藏状态，在更新每个节点的隐藏状态之前，需要先通过公式(1)(2)聚合邻域节点的隐藏状态，然后结合聚合邻域节点得到的信息和上一时步状态信息通过公式(3)更新当前时步的隐藏状态信息，其中时步t为循环次数。

[0056] 5、判别器网络模块训练

[0057] 将生成器得到的16个关键点的热图作为判别器的输入，然后将每个64×64热图处理为256的向量信息表征，此时这16个大小256的向量为关键点的信息表征，这16个关键点可看做图结构中的16个节点，每个节点和邻域节点的关联信息称作边，将节点和边的信息输入到图神经网络得到更新后的节点信息，将更新后的节点信息反复经过图神经网络更新几次得到最终更新完的节点信息，然后将16个大小256的向量处理成为16个大小为1的向量，范围是0到1，从而判别生成器生成的16个关键点质量好坏，增强生成器的性能。

[0058] 6、模型训练

[0059] 整个训练过程为端到端的训练，在训练生成器时，把处理过后的图片数据作为输入，最后一层卷积的输出维度等同于所有的关键点数，得到16个关键点特征。损失函数使用均方差损失函数：

[0060] $$L_{MSE}=\sum_{n=1}^{N} v_n \left\| X_n - Y_n \right\| \qquad (8)$$

[0061] 这里的||·||为欧几里德距离，v为第n个关键点的可见性(0不可见，1可见)，X和Y分别为第n个关键点的预测的结果和真实标签。

[0062] 在训练判别器时，把生成器生成的关键点特征作为输入，最后得到对于16个关键点质量好坏判别的向量。损失函数使用交叉熵损失函数：

[0063] $$L_{BCE}=-\log\left(\frac{\exp(x[gt])}{\sum_i \exp(x[i])}\right) \qquad (9)$$

[0064] 这里的x为预测向量，gt为真实标签(在训练真样本时此处全为1，训练假样本时全为0)。

[0065] 主要损失函数使用均方差损失函数和交叉熵损失函数：

[0066] $L=L_{MSE}+\alpha L_{BCE}$ (10)α为分配的损失权重，得到总的损失函数L。优化器统一般选用Adam优化器来计算梯度并进行反向传播。训练更新参数时需要设置学习率，生成器的初始学习率设置为0.001，判别器的初始学习率比生成器的初始学习率小10倍，然后分别在90和120个epoch时都将学习率降低10倍。每次迭代的图片数设置为32张图片。130个epoch后损失趋于平稳，并在140个epoch时结束训练为。

[0067] 7、模型应用

[0068] 通过上面的训练过程，可以得到多个模型，选取其中最优的模型用于应用测试，图片数据处理在这里并不需要数据增强，只需要把图像调整到256×256大小，然后对数据做归一化即可作为生成器模块的输入。整个的网络模型的参数都固定不动，只要输入图像数据并向前推理即可。在模型应用时不需要判别器模块，只需把生成器最后得到的特征作为预测关键点的特征，然后将预测到的关键点特征解码为坐标点，计算预测坐标点和真实标

签坐标点的欧几里德距离，将此距离做归一化处理即得到预测关键点误差，用于评判模型性能，而预测得到精确的关键点坐标可以作为其他应用，人体动作识别、动画制作、游戏设计以及其他的相关视觉领域等。本发明不仅简化了应用时的模型结构，还减少了参数量，运行速度也极大的提高。

[0069] 以上仅为本发明的较佳实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内所做的任何修改、等同替换和改进等，均在本发明的保护范围之内。

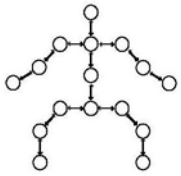


图1

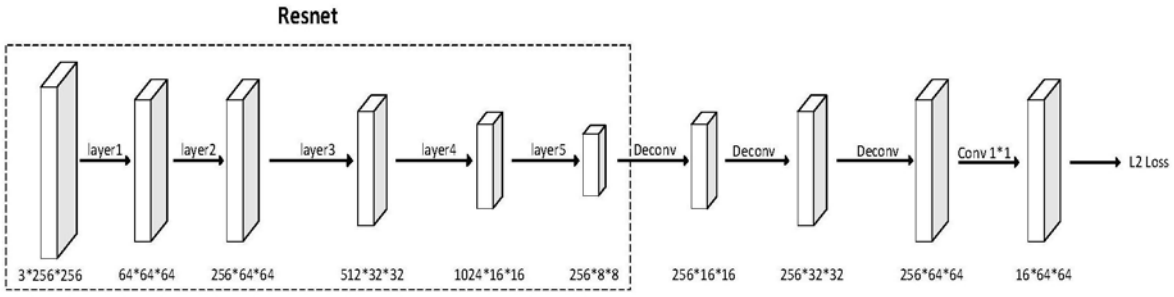

Resnet
layer1
layer2
layer3
layer4
layer5
Deconv
Deconv
Deconv
Conv 1*1
L2 Loss
3*256*256
64*64*64
256*64*64
512*32*32
1024*16*16
256*8*8
256*16*16
256*32*32
256*64*64
16*64*64

图2

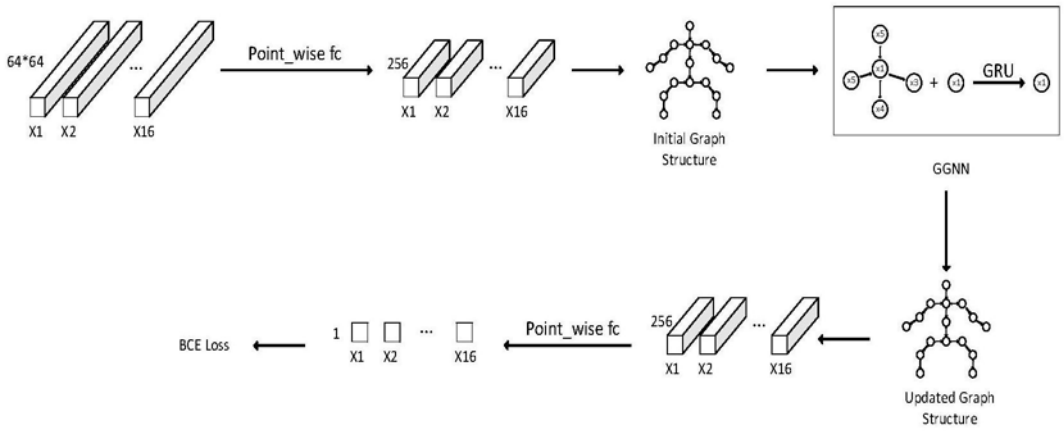

64*64
X1 X2 X16
Point_wise fc
256
X1 X2 X16
Initial Graph Structure
GRU
GGNN
Updated Graph Structure
256
X1 X2 X16
Point_wise fc
1
X1 X2 X16
BCE Loss

图3