(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117635458 A
(43)申请公布日 2024.03.01

(21)申请号 202311659020.5

(22)申请日 2023.12.05

(71)申请人 河南省科学院物理研究所
地址 450046 河南省郑州市金水区明理路266-38号
申请人 河南省科学院

(72)发明人 金贝贝 宋晓辉 李金东 张鹏飞

(74)专利代理机构 深圳市众元信科专利代理有限公司 44757
专利代理师 徐佳辰

(51)Int.Cl.
G06T 5/50(2006.01)
G06V 10/44(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06V 10/62(2022.01)
G06N 3/0475(2023.01)
G06N 3/0464(2023.01)
G06N 3/0442(2023.01)
G06N 3/0455(2023.01)

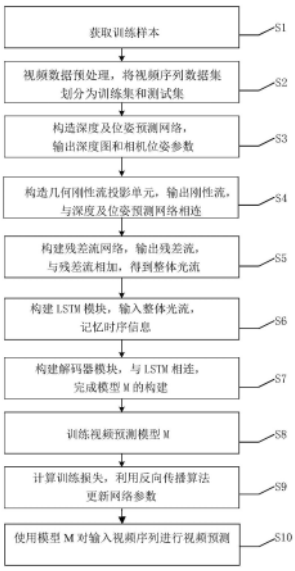
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种基于深度流解析网络的视频预测方法

(57)摘要

本发明公开了一种基于深度流解析网络的视频预测方法,通过将光流解析为刚性流和残差流来预测未来的场景,刚性流表示由于观察者的自我运动而产生的场景动态,残差流对应于场景中其它物体的运动。具体地,本方法提出了一种端到端无监督深度神经网络,通过将场景运动分解为自我运动(相机运动)和以物体为中心的运动来预测未来视频帧。该方法提高了模型解析场景动态信息的能力,具有一定社会价值和现实意义。



CN 117635458 A

1.一种基于差分注意力机制的时空小波分析视频预测方法,其特征在于,包括以下步骤:

S1、获取训练样本;

S2、对视频数据预处理操作;

S3、构造深度及位姿预测网络;

基于卷积神经网络架构,移除原有的全连接层及其后的所有层,仅保留卷积和池化部分,构建深度及位姿预测网络;

S4、构建几何刚性流投影单元,连接到S3中保留的卷积和池化的卷积神经网络架构后面;

S5、构建基于卷积神经网络的残差流网络,输出残差流,与残差流相加,得到整体光流;

S6、构建LSTM模块,输入整体光流,记忆时序信息;

S7、构建解码器模块,连接到S6构建的LSTM网络之后,得到视频预测网络模型M;

S8、训练视频预测模型M;

S9、计算训练损失,利用反向传播算法更新网络参数;

S10、利用训练好的网络对输入的视频序列进行视频帧预测。

一种基于深度流解析网络的视频预测方法

技术领域

[0001] 本发明属于视频分析及预测技术领域,具体涉及一种基于深度流解析网络的视频预测方法。

背景技术

[0002] 基于当前和历史的观察来预测未来情况的能力对机器做出决策至关重要。这项任务对人类来说相对容易,但对机器来说却极具挑战性。近年来,计算机视觉研究人员将注意力集中在视频预测任务上,具体来说,这个任务是指从已经观测的视频帧来预测未来的视频帧。

[0003] 鲁棒有效的视频预测方法不仅需要充分利用空间语义信息,还需要准确掌握时序运动规律。运动动态包含了丰富的场景演化信息,这对于理解环境至关重要,尤其是对于自动驾驶汽车而言。现有的方法几乎都是通过直接光流或帧间差来联合估计背景和前景物体的运动,然而,场景中背景和前景物体的运动是不同源的:前者纯粹来自观察者相机的自我运动,而后者则来自观察者相机的自我运动和物体的残差运动的双重叠加。因此,现有的方法在区分场景静止物体和运动物体方面能力有限,无法高保真地解析场景动态信息。在动态物体密集的复杂城市环境中,这一问题进一步加剧。

[0004] Rushton等人发现,在人类视觉系统中存在一种“流解析机制”,大脑利用其对光流的敏感性将视网膜运动解析为由自我或以物体为中心的运动产生的成分,深度信息在这一过程中也起着重要的作用。首先从观察者的运动对视网膜产生的视觉刺激中估计出自我运动分量,然后从视网膜运动中“减去”自我运动来计算“真实”的以物体为中心的运动估计。这种认知能力帮助人类系统地解决问题和适应新情况。本方法从这种生物“流解析机制”中获得灵感,提出通过场景几何重构来解耦背景变化和以物体为中心的残差运动,从而促进对视频序列中未来帧的推断。

[0005] 已有的视频预测算法可以分为确定性的视频预测方法和随机视频预测方法。确定性视频预测方法的目标是将真实情况与预测结果之间的重建距离最小化。除了确保每帧的预测质量外,还需要提取视频序列中的时序表示。确定性视频预测任务对于自动驾驶、机器人控制等具有重要意义,可以生成足够准确的预测,以做出更安全、更可靠的决策。在确定性方法中,直接像素合成模型试图逐帧直接预测未来的像素强度,它们在特征提取过程中隐式地对场景的动态和静态内容进行建模。Ranzato等人使用k-means对图像块簇中的视频帧进行离散,他们假设非重叠的图像块在k-means离散化空间中是不同的。该方法是基于递归神经网络的模型,在块级进行短期预测,由于整帧是由预测的块组成的,对大型和快速移动的物体的预测是准确的,然而,当涉及到小型和缓慢移动的物体时,仍然有改进的空间。Lotter等人提出了“PredNet”,其灵感来自神经科学的“预测编码”概念。“PredNet”由一系列重复堆叠的模块组成,这些模块试图对模块的输入进行局部预测,尽管表现出一些有希望的结果,但该模型所能预测的时序长度有限。因此,提高长时预测性能成为后续工作的重点。Jin等人利用生成对抗网络来提高预测的真实性。受人类视觉系统的频带分解特性启

发,Jin等人提出利用小波分析探索多频分析实现高保真度和时序一致性的视频预测。Shouno等人提出了一种具有分层结构的深度残差网络来处理大型运动,其中每一层在不同的空间分辨率下对未来状态进行预测。这些不同层的预测通过自上而下的连接合并以生成未来的帧。另一种类型的确定性方法利用变换矩阵的生成来进行视频预测,生成的变换矩阵等价于相邻帧之间的仿射变换。Vondrick等人通过学习转换来处理未来的不确定性和过去的记忆,将过去的记忆与对未来的预测分开。

[0006] 随机视频预测方法认为未来预测是一个多模态任务,它们通常将不确定性编码为潜在变量序列。随机方法通常基于生成对抗网络,变分自编码器等结构。Babaeizadeh等人提出了第一个随机多帧预测的工作,他们提出了一种随机变分视频预测方法,可以预测每个潜在变量样本的不同可能的未来。Denton等人提出了一种随机视频生成模型,该模型结合了确定性帧预测器和随时间变化的随机潜在变量。Lee等人提出第一个通过变分下界和对抗训练来产生高质量预测的工作。

[0007] 虽然已有的视频预测算法已经取得了一定的性能,但它们缺乏对运动信息解耦理解,往往导致预测视频序列模糊和缺乏时序一致性,难以发挥很好的效果。

发明内容

[0008] 本发明实例公开了一种基于深度流解析网络的视频预测方法,通过将光流解析为刚性流和残差流来预测未来的场景,刚性流表示由于观察者的自我运动而产生的场景动态,残差流对应于场景中其它物体的运动。具体地,本方法提出了一种端到端无监督深度神经网络,通过将场景运动分解为自我运动(相机运动)和以物体为中心的运动来预测未来视频帧。该方法提高了模型解析场景动态信息的能力,具有一定社会价值和现实意义。

[0009] 本发明技术方案如下:

[0010] 一种基于深度流解析网络的视频预测方法,包括以下步骤:

[0011] S1、获取训练样本;

[0012] S2、对视频数据预处理操作;

[0013] S3、构造深度及位姿预测网络;

[0014] 基于卷积神经网络架构,移除原有的全连接层及其后的所有层,仅保留卷积和池化部分,构建深度及位姿预测网络;

[0015] S4、构建几何刚性流投影单元,连接到S3中保留的卷积和池化的卷积神经网络架构后面;

[0016] S5、构建基于卷积神经网络的残差流网络,输出残差流,与残差流相加,得到整体光流;

[0017] S6、构建LSTM模块,输入整体光流,记忆时序信息;

[0018] S7、构建解码器模块,连接到S6构建的LSTM网络之后,得到视频预测网络模型M;

[0019] S8、训练视频预测模型M;

[0020] S9、计算训练损失,利用反向传播算法更新网络参数;

[0021] S10、利用训练好的网络对输入的视频序列进行视频帧预测。

[0022] 进一步地,步骤S1具体为:

[0023] 从数据库中获取视频序列数据集,数据集包括针对汽车自动驾驶进行视频预测的

KITTI数据集和Caltech Pedestrain数据集,训练网络时先以其中一个数据集为唯一数据集提取一定数量的视频帧序列作为输入,后续的视频帧为对应的参考结果,随后再以另一个数据集作为唯一数据集进行相同操作。

[0024] 进一步地,步骤S2具体为:

[0025] S21、缩放:将视频帧缩放到原来的θ倍,本实施例中的取值范围为1.0～1.5;

[0026] S22、裁剪:原来的训练样本随机剪切出320*320像素的视频序列;

[0027] S23、HSL调整:对裁剪后样本的色度(Hue)、饱和度(Saturation)和亮度(Lightness)乘以一个随机值δ∈[1.0,1.2],以模拟自然环境的光照变化。

[0028] S24、随后将视频序列数据集划分为训练集和测试集;

[0029] 进一步地,步骤S8具体为:

[0030] 从S1中的输入视频序列提取t帧连续的视频图像序列$X=\{x_1,x_2,\cdots,x_t\}$,将视频图像序列X按顺序输入S7中构建的视频预测网络M提取特征并预测下一个视频帧图像$\hat{x}_{t+1}$。

[0031] 进一步地,步骤S9具体为:

[0032] 将预测的视频帧$\hat{x}_{t+1}$输入到S7的视频预测网络得到预测的$\hat{x}_{t+2}$,如此类推,直到得到要预测的k帧视频序列$\hat{Y}=\{\hat{x}_{t+1},\hat{x}_{t+2},\cdots,\hat{x}_{t+k}\}$;将真实的视频序列$S=\{x_1,x_2,\cdots,x_t,x_{t+1},x_{t+2},\cdots,x_{t+k}\}$与预测的视频帧序列$\hat{S}=\{x_1,x_2,\cdots,x_t,\hat{x}_{t+1},\hat{x}_{t+2},\cdots,\hat{x}_{t+k}\}$对比,计算损失,利用反向传播算法训练网络模型M,训练时所用损失函数分别为:

[0033] $$\mathcal{L}_2(S,\hat{S})=||(S-\hat{S})||_2^2=\sum_{i=t}^{t+k}||(x_i-\hat{x}_i)||_2^2,$$

[0034] $$\mathcal{L}_{GDL}(S,\hat{S})=\sum_{i=1}^{T}\sum_{i,j}||x_{i,j}-x_{i-1,j}|-|\hat{x}_{i,j}-\hat{x}_{i-1,j}||^{\alpha}+||x_{i,j-1}-x_{i,j}|-|\hat{x}_{i,j-1}-\hat{x}_{i,j}||^{\alpha}。$$

[0035] 与现有技术相比,本发明的有益技术效果:

[0036] 1)本发明提出了一种基于深度流解析网络的视频预测方法。在真实场景中,摄像机的自我运动和以物体为中心的运动叠加导致了复杂的动态演化,对动态演变的全面认识和理解是视频预测任务所必需的。先前的研究大多集中在整体运动的处理上,忽略了相机自我运动和物体中心运动的模糊性,导致对整体场景动态的理解不完整。本方法受人类视觉系统的“流解析机制”启发,提出通过场景几何重构来分离背景变化和以物体为中心的残差运动,以方便对视频序列中未来帧的推断。这使得该方法较之于传统视频预测方法能够更好地感知视频中的运动,进而提高了预测的准确性和稳定性。

[0037] 2)本发明强调了在未来预测中消除相机自我运动和物体中心运动的歧义的重要性。将光流解析为与相机运动相关的刚性光流和与物体中心运动相关的残差光流。此外,通过全卷积神经网络从历史帧中同步提取内容信息,通过对内容和运动特征的双重理解,模型取得了更好的预测效果。

[0038] 3)本发明通过引入流解析机制,实现了对视频运动的深刻理解,从而提升模型的准确性和稳定性。因此,本发明在视频预测领域具有重要的应用价值和广阔的发展前景。在实际使用中只需要把视频序列输入生成网络中,通过一次前向传播即可得到结果预测序列,相比传统的视频预测方法有更好的效果。

附图说明

[0039] 图1为本发明视频预测方法流程图；

[0040] 图2为本发明实施过程图；

[0041] 图3为本发明视频预测网络结构示意图。

具体实施方式

[0042] 如图1-3所示,一种基于深度流解析网络的视频预测方法,包括以下步骤：

[0043] S1、获取训练样本

[0044] 从数据库中获取视频序列数据集,数据集包括针对汽车自动驾驶进行视频预测的KITTI数据集和Caltech Pedestrian数据集,训练网络时先以其中一个数据集为唯一数据集提取一定数量的视频帧序列作为输入,后续的视频帧为对应的参考结果,随后再以另一个数据集作为唯一数据集进行相同操作；

[0045] S2、对视频数据预处理操作

[0046] 步骤S2具体为：

[0047] S21、缩放:将视频帧缩放到原来的θ倍,本实施例中的取值范围为1.0～1.5；

[0048] S22、裁剪:原来的训练样本随机剪切出320*320像素的视频序列；

[0049] S23、HSL调整:对裁剪后样本的色度(Hue)、饱和度(Saturation)和亮度(Lightness)乘以一个随机值δ∈[1.0,1.2],以模拟自然环境的光照变化。

[0050] S24、随后将视频序列数据集划分为训练集和测试集；

[0051] S3、构造深度及位姿预测网络；

[0052] 基于卷积神经网络架构,移除原有的全连接层及其后的所有层,仅保留卷积和池化部分,构建深度及位姿预测网络；

[0053] S4、构建几何刚性流投影单元,连接到S3中保留的卷积和池化的卷积神经网络架构后面；

[0054] S5、构建基于卷积神经网络的残差流网络,输出残差流,与残差流相加,得到整体光流；

[0055] S6、构建LSTM模块,输入整体光流,记忆时序信息；

[0056] S7、构建解码器模块,连接到S6构建的LSTM网络之后,得到视频预测网络模型M；

[0057] S8、训练视频预测模型M；

[0058] 步骤S8具体为：

[0059] 从S1中的输入视频序列提取t帧连续的视频图像序列$X=\{x_1,x_2,\cdots,x_t\}$,式中x_i表示第i帧图像,将视频图像序列X按顺序输入S7中构建的视频预测网络M提取特征并预测下一个视频帧图像$\hat{x}_{t+1}$,即t+1时刻的图像帧。

[0060] S9、计算训练损失,利用反向传播算法更新网络参数

[0061] 步骤S9具体为：

[0062] 将预测的视频帧$\hat{x}_{t+1}$输入到S7的视频预测网络得到预测的$\hat{x}_{t+2}$,即t+2时刻的图像帧,如此类推,直到得到要预测的k帧视频序列$\hat{Y}=\{\hat{x}_{t+1},\hat{x}_{t+2},\cdots,\hat{x}_{t+k}\}$;将真实的视频序列$S=\{x_1,x_2,\cdots,x_t,x_{t+1},x_{t+2},\cdots,x_{t+k}\}$与预测的视频帧序列$\hat{S}=\{x_1,x_2,\cdots,x_t,\hat{x}_{t+1},\hat{x}_{t+2},\cdots,\hat{x}_{t+k}\}$对比,

计算损失,利用反向传播算法训练网络模型M,训练时所用损失函数分别为:

[0063] $\mathcal{L}_2(S,\hat{S}) = ||(S-\hat{S})||_2^2 = \sum_{i=t}^{t+k}||(x_i-\hat{x}_i)||_2^2$,

[0064] $\mathcal{L}_{GDL}(S,\hat{S}) = \sum_{i=1}^{T}\sum_{i,j}||x_{i,j}-x_{i-1,j}|-|\hat{x}_{i,j}-\hat{x}_{i-1,j}||^{\alpha} + ||x_{i,j-1}-x_{i,j}|-|\hat{x}_{i,j-1}-\hat{x}_{i,j}||^{\alpha}$。

[0065] S10、利用训练好的网络对输入的视频序列进行视频帧预测。

[0066] 以上所述的实施例仅是对本发明的优选方式进行描述,并非对本发明的范围进行限定,在不脱离本发明设计精神的前提下,本领域普通技术人员对本发明的技术方案做出的各种变形和改进,均应落入本发明权利要求书确定的保护范围内。

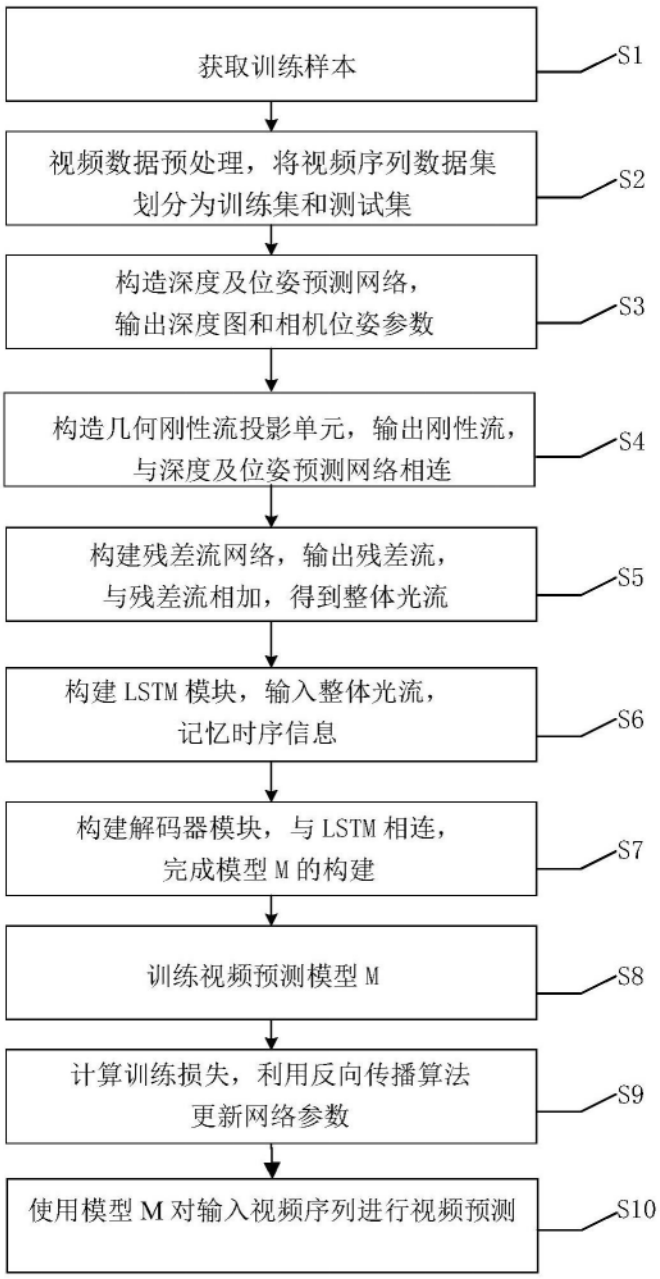

获取训练样本
S1
视频数据预处理，将视频序列数据集划分为训练集和测试集
S2
构造深度及位姿预测网络，输出深度图和相机位姿参数
S3
构造几何刚性流投影单元，输出刚性流，与深度及位姿预测网络相连
S4
构建残差流网络，输出残差流，与残差流相加，得到整体光流
S5
构建 LSTM 模块，输入整体光流，记忆时序信息
S6
构建解码器模块，与 LSTM 相连，完成模型 M 的构建
S7
训练视频预测模型 M
S8
计算训练损失，利用反向传播算法更新网络参数
S9
使用模型 M 对输入视频序列进行视频预测
S10

图1

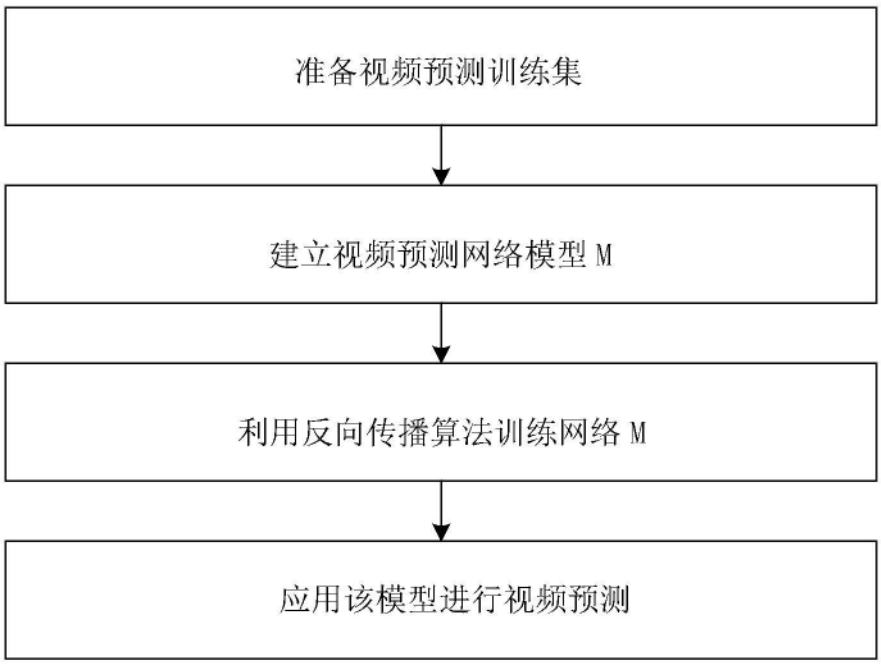

准备视频预测训练集
建立视频预测网络模型 M
利用反向传播算法训练网络 M
应用该模型进行视频预测

图2

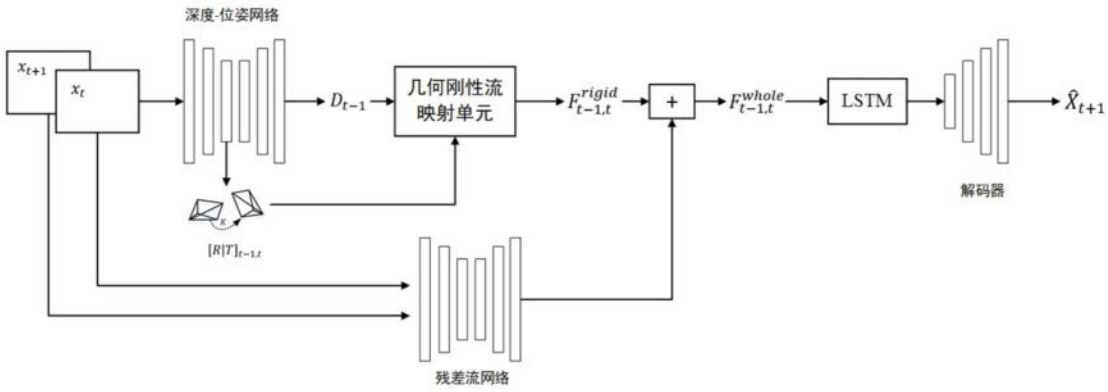

深度-位姿网络
x_{t+1}
x_t
D_{t-1}
几何刚性流
映射单元
$F_{t-1,t}^{rigid}$
+
$F_{t-1,t}^{whole}$
LSTM
$\hat{X}_{t+1}$
解码器
$[R|T]_{t-1,t}$
残差流网络

图3