(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114663496 B
(45)授权公告日 2022.10.18

(21)申请号 202210290482.3
(22)申请日 2022.03.23
(65)同一申请的已公布的文献号
申请公布号 CN 114663496 A
(43)申请公布日 2022.06.24
(73)专利权人 北京科技大学
地址 100083 北京市海淀区学院路30号
专利权人 北京科技大学顺德研究生院
(72)发明人 曾慧 修海鑫 刘红敏 樊彬 张利欣
(74)专利代理机构 北京市广友专利事务所有限责任公司 11237
专利代理师 张仲波
(51)Int.Cl.
G06T 7/70(2017.01)
G06T 7/277(2017.01)
G06T 7/55(2017.01)
G06V 10/80(2022.01)
G06K 9/62(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

(56)对比文件
US 2020041276 A1,2020.02.06
CN 112102399 A,2020.12.18
CN 113108771 A,2021.07.13
CN 110910447 A,2020.03.24
CN 114022527 A,2022.02.08
CN 110490928 A,2019.11.22
US 2022036577 A1,2022.02.03
CN 108665496 A,2018.10.16
YAN WANG et.al.Unsupervised Learning of Accurate Camera Pose and Depth From Video Sequences With Kalman Filter.《IEEE Access》.2019,
张玮奇.基于学习的单目同步定位与地图构建方法研究.《中国优秀博硕士学位论文全文数据库(博士)信息科技辑》.2022,
周凯等.动态环境下融合边缘信息的稠密视觉里程计算法.《哈尔滨工业大学学报》.2021,第53卷(第2期),
(续)

审查员 黄虹

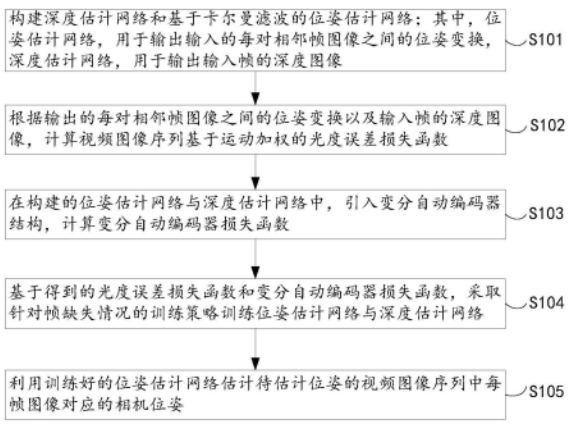
权利要求书4页 说明书13页 附图2页

(54)发明名称

一种基于卡尔曼位姿估计网络的单目视觉里程计方法

(57)摘要

本发明提供一种基于卡尔曼位姿估计网络的单目视觉里程计方法,属于计算机视觉技术领域。所述方法包括:构建深度估计网络和基于卡尔曼滤波的位姿估计网络;根据位姿估计网络输出的每对相邻帧图像之间的位姿变换以及深度估计网络输出的输入帧的深度图像,计算视频图像序列基于运动加权的光度误差损失函数;在构建的位姿估计网络与深度估计网络中,引入变分自动编码器结构,计算变分自动编码器损失函数;基于得到的光度误差损失函数和变分自动编码器损失函数,采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络;利用训练好的位姿估计网络估计每帧图像对应的相机位姿。采用本发明,能够提高相机位姿估计的精度并适应帧缺失的情况。



CN 114663496 B

[转续页]

[接上页]

(56)对比文件

Chunhui Zhao et.al.Pose estimation for multi-camera systems.《2017 IEEE International Conference on Unmanned Systems (ICUS)》.2018,

1.一种基于卡尔曼位姿估计网络的单目视觉里程计方法,其特征在于,包括:

构建深度估计网络和基于卡尔曼滤波的位姿估计网络;其中,位姿估计网络,用于输出输入的每对相邻帧图像之间的位姿变换,深度估计网络,用于输出输入帧的深度图像;

根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像,计算视频图像序列基于运动加权的光度误差损失函数;

在构建的位姿估计网络与深度估计网络中,引入变分自动编码器结构,计算变分自动编码器损失函数;

基于得到的光度误差损失函数和变分自动编码器损失函数,采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络;

利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿;

其中,所述根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像,计算视频图像序列基于运动加权的光度误差损失函数包括:

将位姿估计网络输出的每对相邻帧图像之间的位姿变换相乘得到较长时段的位姿变换,基于得到的较长时段的位姿变换,计算图像之间基于运动加权的光度误差;

根据计算得到的光度误差,计算视频图像序列基于运动加权的光度误差损失函数;

其中,所述将位姿估计网络输出的每对相邻帧图像之间的位姿变换相乘得到较长时段的位姿变换,基于得到的较长时段的位姿变换,计算图像之间基于运动加权的光度误差包括:

对于长度为N的一段视频图像序列,其对应的时刻为$t_0,t_1,\ldots,t_{N-1}$,将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘,得到较长时段的位姿变换$T_{t_j\to t_i}$,其中,$T_{t_j\to t_i}$为由时刻t_j到时刻t_i的图像之间的位姿变换;N为输入位姿估计网络与深度估计网络的每个批次的视频图像序列的长度;

对于图像I_{t_j}上的一个点P_{t_j},其三维坐标由其深度图像$D_{t_j}\left(P_{t_j}\right)$还原;其在图像$I_{t_i}$上对应的投影点$P_{t_i}$表示为:

$$P_{t_i}=KT_{t_j\to t_i}D_{t_j}\left(P_{t_j}\right)K^{-1}P_{t_j}$$

其中,K为摄相机内参数;D_{t_j}为t_j时刻的深度图像;

通过对图像I_{t_i}采样,得到t_j时刻图像I_{t_j}的重构图像$I'_{t_i\to t_j}$:

$$\hat{I}_{t_i\to t_j}\left(P_{t_j}\right)=I_{t_i}\left(P_{t_i}\right)$$

对于P_{t_j}处的像素$I_{t_j}\left(P_{t_j}\right)$,使用$I_{t_i}\left(P_{t_j}\right)$计算其运动加权项$W_{mw}$:

$$W_{mw}=e^{-\left\|I_{t_j}\left(P_{t_j}\right)-I_{t_i}\left(P_{t_j}\right)\right\|_2}$$

利用得到的运动加权项W_{mw},计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差:

$$L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)=\frac{\alpha_0}{2}\left(1-SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)*W_{mw}\right)+\alpha_1\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_1+\alpha_2\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_2$$

其中，$L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$表示图像$I_{t_j}$和$I_{t_i}$之间基于运动加权的光度误差，$SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$表示原图像$I_{t_j}$与重构图像$\hat{I}_{t_i\to t_j}$之间的结构相似性，$\alpha_0$、$\alpha_1$、$\alpha_2$为控制各部分比例的超参数，符号*表示像素间乘积，$||\cdot||_1$表示1范数，$||\cdot||_2$表示2范数。

2.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，所述位姿估计网络包括：位姿测量网络、位姿加权融合网络、位姿更新网络和位姿预测网络；其中，

通过位姿测量网络对输入的相邻帧图像I_{t-1}和I_t进行编码，得到t时刻的位姿测量向量$C_{measure,t}$：

$C_{measure,t}=Measure(I_{t-1},I_t)$

其中，I_{t-1}和I_t分别表示t-1时刻和t时刻的图像，Measure()为所述位姿测量网络；

将位姿测量向量$C_{measure,t}$和位姿预测向量$C_{pred,t}$输入到位姿加权融合网络，得到t时刻的位姿加权融合向量$C_{fuse,t}$：

$C_{fuse,t}=(1-W_t)*C_{measure,t}+W_t*C_{pred,t}$

其中，W_t为位姿加权融合网络中最后一层全连接层输出的[0,1]之间的权重；$C_{pred,t}$为在将相邻帧图像I_{t-2}、I_{t-1}输入位姿估计网络时，位姿预测网络输出的t时刻的位姿预测向量，$C_{pred,t}=Predict(C_{fuse,t-1})$，$C_{fuse,t-1}$为t-1时刻的位姿加权融合向量，Predict为所述位姿预测网络；

将位姿加权融合向量$C_{fuse,t}$输入位姿更新网络估计位姿变换$T_{t\to t-1}$：

$T_{t\to t-1}=Update(C_{fuse,t})$

其中，Update()为所述位姿更新网络；$T_{t\to t-1}$表示从I_{t-1}到I_t的6自由度相对位姿向量，包括：相对旋转和相对位移。

3.根据权利要求2所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，位姿估计网络与深度估计网络都采用编码器-解码器结构。

4.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，在利用得到的运动加权项W_{mw}，计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差之前，所述方法还包括：

确定参与光度误差计算的像素，将其标记为mask：

$$mask=\left[\left\|I_{t_j}-\hat{I}_{t_i\to t_j}\right\|_*<\left\|I_{t_j}-I_{t_i}\right\|_*\right]$$

其中，I_{t_i}为t_i时刻的原图像，I_{t_j}为t_j时刻的原图像，$\hat{I}_{t_i\to t_j}$为从t_i时刻的原图像I_{t_i}采样得到的t_j时刻图像I_{t_j}的重构图像，$||\cdot||_*$代表光度误差，即1范数或2范数；

以便在计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差时，仅用mask标记了的像素进行计算。

5.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，所述光度误差损失函数表示为：

$$L_p = \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} L_p' \left(I_{t_j}, \hat{I}_{t_i \to t_j} \right)$$

其中，L_p表示光度误差损失函数。

6.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，变分自动编码器损失函数表示为：

$$\begin{aligned} L_{VAE} = \quad & \lambda_1 \mathrm{KL}\left(q_d(c_d|x_d) || p_\eta(c) \right) + \\ & \lambda_2 \mathrm{KL}\left(q_p(c_p|x_p) || p_\eta(c) \right) - \\ & \mathbb{E}_{c_d \sim q_d(c_d|x_d), c_p \sim q_p(c_p|x_p)} \left[\log_2 p\left(\hat{x}_d | c_d, c_p \right) \right] \end{aligned}$$

其中，L_{VAE}表示变分自动编码器损失函数，x_d、x_p都表示输入图像，λ_1、λ_2都表示超参数；$p_\eta(c)$为先验分布，c为该分布的自变量；$q_d(c_d|x_d)$为深度估计网络编码c_d的被采样分布；$q_p(c_p|x_p)$为深度估计网络编码c_p的被采样分布，KL(•)为KL散度，$\mathrm{KL}(q_d(c_d|x_d)||p_\eta(c))$表示$q_d(c_d|x_d)$对于$p_\eta(c)$的KL散度，$\mathrm{KL}(q_p(c_p|x_p)||p_\eta(c))$表示$q_p(c_p|x_p)$对于$p_\eta(c)$的KL散度，$p\left(\hat{x}_d|c_d, c_p\right)$为将$c_d$与$c_p$分别输入深度估计网络与位姿估计网络的解码器得到的输出，进而生成的重构图像$\hat{x}_d$的概率分布，$\mathbb{E}[\cdot]$表示数学期望，$c_d \sim q_d(c_d|x_d)$表示c_d服从$q_d(c_d|x_d)$，$c_p \sim q_p(c_p|x_p)$表示c_p服从$q_p(c_p|x_p)$，$\mathbb{E}_{c_d \sim q_d(c_d|x_d), c_p \sim q_p(c_p|x_p)} \left[\log_2 p\left(\hat{x}_d | c_d, c_p \right) \right]$表示在满足$c_d \sim q_d(c_d|x_d)$及$c_p \sim q_p(c_p|x_p)$的条件下，$\log_2 p\left(\hat{x}_d | c_d, c_p \right)$的数学期望；$c_d \sim q_d(c_d|x_d)$表示$c_d$服从$q_d(c_d|x_d)$分布；$c_p \sim q_p(c_p|x_p)$表示$c_p$服从$q_p(c_p|x_p)$分布。

7.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，其特征在于，所述基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络包括：

对于深度估计网络的输出，计算深度平滑损失函数：

$$L_s = |\partial_x D_t^*| e^{-|\partial_x I_t|} + |\partial_y D_t^*| e^{-|\partial_y I_t|}$$

其中，$D_t^* = \frac{1}{D_t}$为视差，与深度图像D_t成反比例关系，∂_x、∂_y分别表示x方向与y方向上的偏导数，I_t为t时刻的图像；

基于得到的深度平滑损失函数、光度误差损失函数和变分自动编码器损失函数，确定最终的损失函数L：

$L = L_p + \lambda L_s + L_{VAE}$

其中，λ为控制深度平滑损失函数比例的超参数，L_p表示光度误差损失函数，L_{VAE}表示变分自动编码器损失函数；

利用得到的最终的损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深

度估计网络。

8.根据权利要求1所述的基于卡尔曼位姿估计网络的单目视觉里程计方法,其特征在于,所述采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络包括:

将一个批次的视频图像序列中的所有图像输入到位姿估计网络与深度估计网络中,同时对位姿估计网络与深度估计网络进行训练;

将一个批次的视频图像序列中的所有图像输入到深度估计网络中,并将该批次的视频图像序列中的一帧或多帧图像置零后再输入位姿估计网络,对位姿估计网络与深度估计网络再进行训练。

一种基于卡尔曼位姿估计网络的单目视觉里程计方法

技术领域

[0001] 本发明涉及计算机视觉技术领域，特别是指一种基于卡尔曼位姿估计网络的单目视觉里程计方法。

背景技术

[0002] 视觉里程计作为同时定位与建图技术的一部分，广泛应用在机器人导航、自动驾驶、增强现实、可穿戴计算等领域。视觉里程计是指根据输入视频图像帧估计相机当前的位置与姿态的方法。根据采用传感器的种类和数目不同，视觉里程计可分为单目视觉里程计、双目视觉里程计以及融合惯性信息的视觉里程计等。其中，单目视觉里程计具有着仅需要一个相机，对硬件要求较低、无需矫正等优点。

[0003] 传统的视觉里程计方法首先进行图像特征提取与匹配，然后根据几何关系估计相邻两帧之间的相对位姿。这种方法在实际应用中取得了不错的结果，是当前视觉里程计的主流方法，但其存在计算性能与鲁棒性难以平衡的问题。

[0004] 基于深度学习的单目视觉里程计可分为有监督的方法和自监督的方法。自监督的方法仅仅需要输入视频图像帧，不需要采集真实的位姿，没有对额外设备的依赖，适用性比有监督的方法更为广泛。

[0005] 现有的许多自监督方法没有考虑帧与帧之间的关联，帧间的信息没有被充分利用，导致训练出的网络难以估计出更为精确的位姿，也不能适应帧缺失的情况。此外，场景中的运动物体，其与场景的欧氏变换不一致，不满足静态场景的假设，难以用一个欧氏变换去描述场景的运动，导致网络的估计结果出现偏差。

发明内容

[0006] 本发明实施例提供了一种基于卡尔曼位姿估计网络的单目视觉里程计方法，能够提高相机位姿估计的精度并适应帧缺失的情况。所述技术方案如下：

[0007] 本发明实施例提供了一种基于卡尔曼位姿估计网络的单目视觉里程计方法，包括：

[0008] 构建深度估计网络和基于卡尔曼滤波的位姿估计网络；其中，位姿估计网络，用于输出输入的每对相邻帧图像之间的位姿变换，深度估计网络，用于输出输入帧的深度图像；

[0009] 根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像，计算视频图像序列基于运动加权的光度误差损失函数；

[0010] 在构建的位姿估计网络与深度估计网络中，引入变分自动编码器结构，计算变分自动编码器损失函数；

[0011] 基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络；

[0012] 利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0013] 进一步地，所述位姿估计网络包括：位姿测量网络、位姿加权融合网络、位姿更新网络和位姿预测网络；其中，

[0014] 通过位姿测量网络对输入的相邻帧图像I_{t-1}和I_t进行编码，得到t时刻的位姿测量向量$C_{measure,t}$：

[0015] $C_{measure,t}=Measure(I_{t-1},I_t)$

[0016] 其中，I_{t-1}和I_t分别表示t-1时刻和t时刻的图像，Measure()为所述位姿测量网络；

[0017] 将位姿测量向量$C_{measure,t}$和位姿预测向量$C_{pred,t}$输入到位姿加权融合网络，得到t时刻的位姿加权融合向量$C_{fuse,t}$：

[0018] $C_{fuse,t}=(1-W_t)*C_{measure,t}+W_t*C_{pred,t}$

[0019] 其中，W_t为位姿加权融合网络中最后一层全连接层输出的[0,1]之间的权重；$C_{pred,t}$为在将相邻帧图像I_{t-2}、I_{t-1}输入位姿估计网络时，位姿预测网络输出的t时刻的位姿预测向量，$C_{pred,t}=Predict(C_{fuse,t-1})$，$C_{fuse,t-1}$为t-1时刻的位姿加权融合向量，Predict为所述位姿预测网络；

[0020] 将位姿加权融合向量$C_{fuse,t}$输入位姿更新网络估计位姿变换$T_{t\to t-1}$：

[0021] $T_{t\to t-1}=Update(C_{fuse,t})$

[0022] 其中，Update()为所述位姿更新网络；$T_{t\to t-1}$表示从I_{t-1}到I_t的6自由度相对位姿向量，包括：相对旋转和相对位移。

[0023] 进一步地，位姿估计网络与深度估计网络都采用编码器-解码器结构。

[0024] 进一步地，所述根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像，计算视频图像序列基于运动加权的光度误差损失函数包括：

[0025] 将位姿估计网络输出的每对相邻帧图像之间的位姿变换相乘得到较长时段的位姿变换，基于得到的较长时段的位姿变换，计算图像之间基于运动加权的光度误差；

[0026] 根据计算得到的光度误差，计算视频图像序列基于运动加权的光度误差损失函数。

[0027] 进一步地，所述将位姿估计网络输出的每对相邻帧图像之间的位姿变换相乘得到较长时段的位姿变换，基于得到的较长时段的位姿变换，计算图像之间基于运动加权的光度误差包括：

[0028] 对于长度为N的一段视频图像序列，其对应的时刻为$t_0,t_1,...,t_{N-1}$，将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘，得到较长时段的位姿变换$T_{t_j\to t_i}$，其中，$T_{t_j\to t_i}$为由时刻t_j到时刻t_i的图像之间的位姿变换；N为输入位姿估计网络与深度估计网络的每个批次的视频图像序列的长度；

[0029] 对于图像I_{t_j}上的一个点P_{t_j}，其三维坐标由其深度图像$D_{t_j}\left(P_{t_j}\right)$还原；其在图像$I_{t_i}$上对应的投影点$P_{t_i}$表示为：

[0030] $$P_{t_i}=KT_{t_j\to t_i}D_{t_j}\left(P_{t_j}\right)K^{-1}P_{t_j}$$

[0031] 其中，K为摄相机内参数；D_{t_j}为t_j时刻的深度图像；

[0032] 通过对图像I_{t_i}采样，得到t_j时刻图像I_{t_j}的重构图像$I'_{t_i\to t_j}$：

[0033] $$\hat{I}_{t_i\to t_j}\left(P_{t_j}\right)=I_{t_i}\left(P_{t_i}\right)$$

[0034] 对于P_{t_j}处的像素$I_{t_j}\left(P_{t_j}\right)$，使用$I_{t_i}\left(P_{t_j}\right)$计算其运动加权项$W_{mw}$：

[0035] $$W_{mw}=e^{-\left\|I_{t_j}\left(P_{t_j}\right)-I_{t_i}\left(P_{t_j}\right)\right\|_2}$$

[0036] 利用得到的运动加权项W_{mw}，计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差：

[0037] $$\begin{aligned}L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)=\ &\frac{\alpha_0}{2}\left(1-SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)*W_{mw}\right)+\\&\alpha_1\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_1+\\&\alpha_2\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_2\end{aligned}$$

[0038] 其中，$L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$表示图像$I_{t_j}$和$I_{t_i}$之间基于运动加权的光度误差，$SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$表示原图像$I_{t_j}$与重构图像$\hat{I}_{t_i\to t_j}$之间的结构相似性，$\alpha_0$、$\alpha_1$、$\alpha_2$为控制各部分比例的超参数，符号*表示像素间乘积，$\|\cdot\|_1$表示1范数，$\|\cdot\|_2$表示2范数。

[0039] 进一步地，在利用得到的运动加权项W_{mw}，计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差之前，所述方法还包括：

[0040] 确定参与光度误差计算的像素，将其标记为mask：

[0041] $$mask=\left[\left\|I_{t_j}-\hat{I}_{t_i\to t_j}\right\|_*<\left\|I_{t_j}-I_{t_i}\right\|_*\right]$$

[0042] 其中，I_{t_i}为t_i时刻的原图像，I_{t_j}为t_j时刻的原图像，$\hat{I}_{t_i\to t_j}$为从t_i时刻的原图像I_{t_i}采样得到的t_j时刻图像I_{t_j}的重构图像，$\|\cdot\|_*$代表光度误差，即1范数或2范数；

[0043] 以便在计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差时，仅用mask标记了的像素进行计算。

[0044] 进一步地，所述光度误差损失函数表示为：

[0045] $$L_p=\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$$

[0046] 其中，L_p表示光度误差损失函数。

[0047] 进一步地，变分自动编码器损失函数表示为：

[0048] $$\begin{aligned}L_{VAE}=\ &\lambda_1\mathrm{KL}\left(q_d(c_d|x_d)||p_\eta(c)\right)+\\&\lambda_2\mathrm{KL}\left(q_p(c_p|x_p)||p_\eta(c)\right)-\\&\mathbb{E}_{c_d\sim q_d(c_d|x_d),c_p\sim q_p(c_p|x_p)}\left[\log_2 p\left(\hat{x}_d|c_d,c_p\right)\right]\end{aligned}$$

[0049] 其中，L_{VAE}表示变分自动编码器损失函数，x_d、x_p都表示输入图像，λ_1、λ_2都表示超参数；$p_\eta(c)$为先验分布，c为该分布的自变量；$q_d(c_d|x_d)$为深度估计网络编码c_d的被采样分布；$q_p(c_p|x_p)$为深度估计网络编码c_p的被采样分布，KL(•)为KL散度，$KL(q_d(c_d|x_d)||p_\eta(c))$表示$q_d(c_d|x_d)$对于$p_\eta(c)$的KL散度，$KL(q_p(c_p|x_p)||p_\eta(c))$表示$q_p(c_p|x_p)$对于$p_\eta(c)$的KL散度，$p(\hat{x}_d|c_d,c_p)$为将$c_d$与$c_p$分别输入深度估计网络与位姿估计网络的解码器得到的输出，进而生成的重构图像$\hat{x}_d$的概率分布，$\mathbb{E}[\cdot]$表示数学期望，$c_d\sim q_d(c_d|x_d)$表示c_d服从$q_d(c_d|x_d)$，$c_p\sim q_p(c_p|x_p)$表示c_p服从$q_p(c_p|x_p)$，$\mathbb{E}_{c_d\sim q_d(c_d|x_d),c_p\sim q_p(c_p|x_p)}[\log_2 p(\hat{x}_d|c_d,c_p)]$表示在满足$c_d\sim q_d(c_d|x_d)$及$c_p\sim q_p(c_p|x_p)$的条件下，$\log_2 p(\hat{x}_d|c_d,c_p)$的数学期望；$c_d\sim q_d(c_d|x_d)$表示$c_d$服从$q_d(c_d|x_d)$分布；$c_p\sim q_p(c_p|x_p)$表示$c_p$服从$q_p(c_p|x_p)$分布。

[0050] 进一步地，所述基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络包括：

[0051] 对于深度估计网络的输出，计算深度平滑损失函数：

[0052] $$L_s = |\partial_x D_t^*|e^{-|\partial_x I_t|} + |\partial_y D_t^*|e^{-|\partial_y I_t|}$$

[0053] 其中，$D_t^* = \frac{1}{D_t}$为视差，与深度图像D_t成反比例关系，∂_x、∂_y分别表示x方向与y方向上的偏导数，I_t为t时刻的图像；

[0054] 基于得到的深度平滑损失函数、光度误差损失函数和变分自动编码器损失函数，确定最终的损失函数L：

[0055] $L=L_p+\lambda L_s+L_{VAE}$

[0056] 其中，λ为控制深度平滑损失函数比例的超参数，L_p表示光度误差损失函数，L_{VAE}表示变分自动编码器损失函数；

[0057] 利用得到的最终的损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络。

[0058] 进一步地，所述采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络包括：

[0059] 将一个批次的视频图像序列中的所有图像输入到位姿估计网络与深度估计网络中，对位姿估计网络与深度估计网络进行训练；

[0060] 将一个批次的视频图像序列中的所有图像输入到深度估计网络中，并将该批次的视频图像序列中的一帧或多帧图像置零后再输入位姿估计网络，对位姿估计网络与深度估计网络再进行训练。

[0061] 本发明实施例所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，至少具有以下优点：

[0062] (1)针对现有的许多自监督方法没有考虑帧与帧之间的关联，帧间的信息没有被充分利用，导致训练出的网络难以估计出更为精确的位姿，也不能适应帧缺失的情况的问题，本实施例构建了基于卡尔曼滤波的位姿估计网络，并以此为基础，设计了针对帧缺失情况的训练策略，使得位姿估计网络可以利用帧间的信息估计当前的位姿，更加适应帧缺失的情况；

[0063] (2)针对场景中可能存在的运动物体与场景的欧氏变换不一致,不满足静态场景的假设,难以用一个欧氏变换去描述场景的运动,导致位姿估计网络的估计结果出现偏差的问题,本实施例采用一种运动加权策略,同时在位姿估计网络与深度估计网络中引入变分自动编码器结构,使得位姿估计网络与深度估计网络可以在训练阶段更关注场景中的静止物体,提高网络泛化能力,提升网络性能。

附图说明

[0064] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0065] 图1为本发明实施例提供的基于卡尔曼位姿估计网络的单目视觉里程计方法的流程示意图;

[0066] 图2为本发明实施例提供的位姿估计网络的结构示意图;

[0067] 图3为本发明实施例提供的基于卡尔曼位姿估计网络的单目视觉里程计方法的工作流程示意图;

[0068] 图4为本发明实施例提供的方法在KITTI里程计数据集中序列09、10上估计的轨迹示意图。

具体实施方式

[0069] 为使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明实施方式作进一步地详细描述。

[0070] 如图1所示,本发明实施例提供了一种基于卡尔曼位姿估计网络的单目视觉里程计方法,包括:

[0071] S101,构建深度估计网络(DepthNet)和基于卡尔曼滤波的位姿估计网络(KF-PoseNet);其中,位姿估计网络,用于输出输入的每对相邻帧图像之间的位姿变换,深度估计网络,用于输出输入帧的深度图像;

[0072] 如图2所示,所述位姿估计网络包括:位姿测量网络、位姿加权融合网络、位姿更新网络和位姿预测网络;其中,如表1所示,

[0073] 位姿测量网络包括ResNet50、三层卷积层以及一个全局平均池化层;三层卷积层中的前两层卷积层以ReLU(Rectification Linear Unit,整流线性单元)为激活函数,最后一层卷积层为纯卷积层,无激活函数;位姿测量网络的输入经过ResNet50后,再依次经过三层卷积层,最后经过全剧平均池化层输出;位姿测量网络使用ResNet50结构作为编码器;

[0074] 位姿加权融合网络包括4个全连接层和一个加权融合层;4个全连接层中前三层全连接层使用ReLU作为激活函数,最后一层全连接层使用Sigmoid函数作为激活函数;$C_{measure,t}$和$C_{pred,t}$输入第一个全连接层后,依次经过后三个全连接层,输出取值范围为0-1的权重系数;该权重系数进一步与$C_{measure,t}$和$C_{pred,t}$一同送入加权融合层;

[0075] 位姿更新网络包含4个全连接层,前三个全连接层使用ReLU作为激活函数;4个全连接层依次相连;

[0076] 类似于位姿更新网络，位姿预测网络同样包含4个全连接层，4个全连接层依次相连。

[0077] 表1 KF-PoseNet网络结构

[0078]

module	layer	inplane	kernel	outplane
位姿测量网络	resnet50	6		2048
	conv+relu	2048	1	256
	conv+relu	256	3	256
	conv	256	3	256
位姿预测网络	fc+relu	256		1024
	fc+relu	1024		1024
	fc+relu	1024		1024
	fc	1024		256
位姿加权融合网络	fc+relu	512		256
	fc+relu	256		256
	fc+relu	256		256
	fc+sigmoid	256		1
位姿更新网络	fc+relu	256		256
	fc+relu	256		256
	fc+relu	256		256
	fc	256		6

[0079] 本实施例中，所述位姿估计网络的工作流程为：

[0080] 通过位姿测量网络对输入的相邻帧图像I_{t-1}和I_t进行编码，得到t时刻的位姿测量向量$C_{measure,t}$：

[0081] $C_{measure,t}=Measure(I_{t-1},I_t)$

[0082] 其中，I_{t-1}和I_t分别表示t-1时刻和t时刻的图像，Measure()为所述位姿测量网络；应当注意的是，$C_{measure,t}$并不是6自由度位姿向量，而仅仅是包含图像对(I_{t-1},I_t)位姿信息的编码向量；

[0083] 将位姿测量向量$C_{measure,t}$和位姿预测向量$C_{pred,t}$输入到位姿加权融合网络，得到t时刻的位姿加权融合向量$C_{fuse,t}$：

[0084] $C_{fuse,t}=(1-W_t)*C_{measure,t}+W_t*C_{pred,t}$

[0085] 其中，$W_t=Weight(C_{measure,t},C_{pred,t})$为位姿加权融合网络中最后一层全连接层输出的[0,1]之间的权重，Weight为所述位姿加权融合网络中的4个全连接层；$C_{pred,t}$为在将相邻帧图像I_{t-2}、I_{t-1}输入位姿估计网络时，位姿预测网络输出的t时刻的位姿预测向量，$C_{pred,t}=Predict(C_{fuse,t-1})$，$C_{fuse,t-1}$为t-1时刻的位姿加权融合向量，Predict为所述位姿预测网络；

[0086] 将位姿加权融合向量$C_{fuse,t}$输入位姿更新网络估计最终的位姿变换$T_{t\to t-1}$：

[0087] $T_{t\to t-1}=Update(C_{fuse,t})$

[0088] 其中，Update()为所述位姿更新网络；$T_{t\to t-1}$表示从I_{t-1}到I_t的6自由度相对位姿向量。

[0089] 如图3所示，KF-PoseNet的输入为相邻两帧图像，输出为6自由度相对位姿向量，其前三个元素表示3自由度相对旋转R，后三个元素表示3自由度相对位移t。

[0090] 本实施例中，位姿估计网络与深度估计网络都采用编码器-解码器结构，所述位姿

估计网络中的编码器为位姿测量网络中的ResNet50结构，所述位姿估计网络的解码器为位姿测量网络中除ResNet50以外其余的结构、位姿加权融合网络、位姿预测网络和位姿更新网络。

[0091] 本实施例中，深度估计网络(DepthNet)同样选择ResNet50结构作为编码器，以类似于DispNet解码器的多层反卷积结构作为解码器，并通过跳跃链接结构与编码器连接，输出层激活函数为Sigmoid。本实施例中，DepthNet的输入为单帧图像，输出为归一化的视差D*。要获得深度D，需要对获得的视差取倒数D＝1/(aD*+b)，其中，a和b为限制输出取值范围的参数，使输出深度为0.1到100之间。

[0092] 本实施例中，为了控制内存占用的同时尽可能地保留细节，将位姿估计网络与深度估计网络的输入RGB图像缩放为了832×256的大小。

[0093] 在本实施例中，设所述一对相邻帧图像为当前时刻t的图像I_t与上一时刻t-1的图像I_{t-1}。将相邻帧图像I_t和I_{t-1}输入构建的位姿估计网络与深度估计网络中，得到所述相邻帧图像之间的位姿变换$T_{t\to t-1}$，与每个输入帧的深度图像Dt。

[0094] S102，根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像，计算视频图像序列基于运动加权的光度误差损失函数；具体可以包括以下步骤：

[0095] A1，将位姿估计网络输出的每对相邻帧图像之间的位姿变换相乘得到较长时段的位姿变换，基于得到的较长时段的位姿变换，计算图像之间基于运动加权的光度误差；

[0096] 本实施例中，一个场景中往往还可能存在一些快速移动的运动物体。这些物体与相机的欧氏变换不一致。显然将这些物体对应的像素在训练网络时同等对待是不合理的。对于数据集中运动幅度不大、光照变化不明显的情况，相邻两帧中在同一位置的像素的亮度往往不会有太大变化。基于这一点，为了降低快速移动物体的影响，本发明设计了基于运动加权的光度误差。且为了使网络能考虑较长时间上的位姿变换的一致性，本实施例在计算基于运动加权的光度误差时利用了连续多帧图像计算了长时间位姿约束的光度误差，具体的：

[0097] 对于长度为N的一段视频图像序列，其对应的时刻为$t_0, t_1, \ldots, t_{N-1}$，将位姿估计网络输出的每对相邻帧图像之间的位姿进行累积相乘，得到较长时段的位姿变换$T_{t_j\to t_i}$：

[0098] $$T_{t_j\to t_i} = T_{t_{i+1}\to t_i} T_{t_{i+2}\to t_{i+1}} \cdots T_{t_{j-1}\to t_j}$$

[0099] 其中，$T_{t_j\to t_i}$为由时刻t_j到时刻t_i的图像之间的位姿变换；N为输入位姿估计网络与深度估计网络的每个批次的视频图像序列的长度；

[0100] 接着，对于图像I_{t_j}上的一个点P_{t_j}，其三维坐标可以由其深度图像$D_{t_j}\left(P_{t_j}\right)$还原；则其在图像$I_{t_i}$上对应的投影点$P_{t_i}$可以由如下公式计算获得：

[0101] $$P_{t_i} = K T_{t_j\to t_i} D_{t_j}\left(P_{t_j}\right) K^{-1} P_{t_j}$$

[0102] 其中，K为摄相机内参数；D_{t_j}为t_j时刻的深度图像；

[0103] 上述公式忽略了部分齐次坐标系的计算；

[0104] 通过对图像I_{t_i}采样，得到t_j时刻图像I_{t_j}的重构图像$I'_{t_i\to t_j}$：

[0105] $$\hat{I}_{t_i\to t_j}\left(P_{t_j}\right)=I_{t_i}\left(P_{t_i}\right)$$

[0106] 然后，对于P_{t_j}处的像素$I_{t_j}\left(P_{t_j}\right)$，可以使用$I_{t_i}\left(P_{t_j}\right)$计算其运动加权项$W_{mw}$：

[0107] $$W_{mw}=e^{-\left\|I_{t_j}\left(P_{t_j}\right)-I_{t_i}\left(P_{t_j}\right)\right\|_2}$$

[0108] 最后，利用得到的运动加权项W_{mw}，计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差$L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$：

[0109] $$L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)=\frac{\alpha_0}{2}\left(1-SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)*W_{mw}\right)+\alpha_1\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_1+\alpha_2\left\|\left(I_{t_j}-\hat{I}_{t_i\to t_j}\right)*W_{mw}\right\|_2$$

[0110] 其中，$SSIM\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$表示原图像$I_{t_j}$与重构图像$\hat{I}_{t_i\to t_j}$之间的结构相似性，$\alpha_0$、$\alpha_1$、$\alpha_2$为控制各部分比例的超参数，符号*表示像素间乘积，$\|\cdot\|_1$表示1范数，$\|\cdot\|_2$表示2范数。

[0111] 本实施例中，使用上述的运动加权项W_{mw}对所计算的广度误差逐像素加权，得到所述运动加权的光度误差。

[0112] 进一步地，考虑到当视野中存在相对相机静止的物体时，可能会影响深度估计的精确度，导致估计的深度变为无穷大。为此，本实施例中还使用一种自动标记静止像素的方法，并从训练过程中将之移除。具体而言，把当前图像与参考图像之间的误差小于重构误差的像素看作相对于相机静止的像素，仅利用重构误差小于当前图像与参考图像之间的误差的像素(即参与光度误差计算的像素)训练深度网络。

[0113] 本实施例中，确定参与光度误差计算的像素，将其标记为mask：

[0114] $$mask=\left[\|I_{t_j}-\hat{I}_{t_i\to t_j}\|_*<\|I_{t_j}-I_{t_i}\|_*\right]$$

[0115] 其中，I_{t_i}为t_i时刻的原图像，I_{t_j}为t_j时刻的原图像，$\hat{I}_{t_i\to t_j}$为从t_i时刻的原图像I_{t_i}采样得到的t_j时刻图像I_{t_j}的重构图像，$\|\cdot\|_*$代表光度误差，即1范数或2范数；

[0116] 以便在计算图像I_{t_j}和I_{t_i}之间基于运动加权的光度误差时，仅用mask标记了的像素进行计算，进而使用mask标记了的像素进行网络训练。

[0117] A2，根据计算得到的光度误差，计算视频图像序列运动加权的光度误差损失函数L_p：$L_p=\sum_{i=0}^{N-1}\sum_{j=i+1}^{N}L_p'\left(I_{t_j},\hat{I}_{t_i\to t_j}\right)$

[0118] 其中，L_p'表示所述运动加权的光度误差。

[0119] S103，在构建的位姿估计网络与深度估计网络中，引入变分自动编码器结构，计算

变分自动编码器损失函数；

[0120] 在本实施例中，KF-PoseNet与DepthNet都采用编码器-解码器结构；为了提高解码器的输出对其输入的编码中的噪声的鲁棒性，提高网络的泛化能力，将变分自动编码器(Variational Auto-Encoder，VAE)结构引入到了KF-PoseNet和DepthNet中；

[0121] 以深度估计网络为例；

[0122] 深度估计网络的编码器将输入图像$x_d=I_t$映射到编码空间，得到均值向量$E_d(x_d)$；

[0123] 进一步地，设$q_d(c_d|x_d)$为待输入到解码器的编码c_d的被采样分布，将其设为均值为输入图像的均值E_d，协方差为输入图像的协方差Σ_d的高斯分布$\mathcal{N}(c_d|E_d,\Sigma_d)$；在$q_d(c_d|x_d)$分布中随机采样得到编码$c_d$，其中$c_d$服从$q_d(c_d|x_d)$分布，用$c_d\sim q_d(c_d|x_d)$表示；

[0124] 进一步地，将编码c_d输入解码器得到输入图像的深度图像；

[0125] 为了满足深度网络反向传播的需要，本实施例中，在编码空间对编码进行随机采样时，采用如下重参数化方法，将随机采样过程变为可微操作：令η为服从零均值单位协方差高斯分布$\mathcal{N}(\eta|0,I)$的随机向量：$\eta\sim\mathcal{N}(\eta|0,I)$，其中I为单位矩阵，则对$c_d\sim q_d(c_d|x_d)$的采样操作可以通过$c_d=E_d(x_d)+\Sigma_d\eta$实现，其中$\Sigma_d$为输入图像的协方差；

[0126] 位姿估计网络同理；

[0127] 进一步地，计算VAE损失函数L_{VAE}为：

[0128]
$$\begin{aligned}L_{VAE}=\ & \lambda_1\mathrm{KL}\left(q_d(c_d|x_d)||p_\eta(c)\right)+\\ & \lambda_2\mathrm{KL}\left(q_p(c_p|x_p)||p_\eta(c)\right)-\\ & \mathbb{E}_{c_d\sim q_d(c_d|x_d),c_p\sim q_p(c_p|x_p)}\left[\log_2 p(\hat{x}_d|c_d,c_p)\right]\end{aligned}$$

[0129] 其中，x_d、x_p都表示输入图像，超参数λ_1、λ_2用于控制目标项的权重，$p_\eta(c)$为先验分布，c为该分布的自变量；$q_d(c_d|x_d)$为深度估计网络编码c_d的被采样分布，$q_p(c_p|x_p)$为深度估计网络编码c_p的被采样分布，KL(•)为KL散度，$\mathrm{KL}(q_d(c_d|x_d)||p_\eta(c))$表示$q_d(c_d|x_d)$对于$p_\eta(c)$的KL散度，$\mathrm{KL}(q_p(c_p|x_p)||p_\eta(c))$表示$q_p(c_p|x_p)$对于$p_\eta(c)$的KL散度，$p(\hat{x}_d|c_d,c_p)$为将$c_d$与$c_p$分别输入深度估计网络与位姿估计网络的解码器得到的输出，进而生成的重构图像$\hat{x}_d$的概率分布，$\mathbb{E}[\cdot]$表示数学期望，$c_d\sim q_d(c_d|x_d)$表示c_d服从$q_d(c_d|x_d)$，$c_p\sim q_p(c_p|x_p)$表示c_p服从$q_p(c_p|x_p)$，$\mathbb{E}_{c_d\sim q_d(c_d|x_d),c_p\sim q_p(c_p|x_p)}\left[\log_2 p(\hat{x}_d|c_d,c_p)\right]$表示在满足$c_d\sim q_d(c_d|x_d)$及$c_p\sim q_p(c_p|x_p)$的条件下，$\log_2 p(\hat{x}_d|c_d,c_p)$的数学期望；公式中前两项控制KL散度惩罚隐藏编码的分布背离先验分布的倾向；最后一项最小化非负对数似然项，等价于最小化光度误差损失函数；因此，VAE损失函数实际仅为公式中的前两项。

[0130] 本实施例中，先验分布$p_\eta(c)$为0均值的高斯分布$p_\eta(c)=\mathcal{N}(c|0,I)$。

[0131] S104，基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络；具体可以包括以下步骤：

[0132] 首先，考虑到在三维空间中的一个纹理稳定平面内，其在深度图像中的深度往往不会产生太剧烈的变化。因此，在本实施例中，对于深度估计网络的输出，还按如下公式计

算深度平滑损失函数L_s：

[0133] $$L_s = |\partial_x D_t^*|e^{-|\partial_x I_t|} + |\partial_y D_t^*|e^{-|\partial_y I_t|}$$

[0134] 其中，$D_t^* = \frac{1}{D_t}$为视差，与深度图像D_t成反比例关系，∂_x、∂_y分别表示x方向与y方向上的偏导数，I_t为t时刻的图像；

[0135] 在本实施例中，上述深度平滑损失函数对每个批次中的每帧图像都进行了计算；

[0136] 接着，基于得到的深度平滑损失函数、光度误差损失函数和变分自动编码器损失函数，确定最终的损失函数L：

[0137] $L=L_p+\lambda L_s+L_{VAE}$

[0138] 其中，λ为控制深度平滑损失函数比例的超参数，L_p表示光度误差损失函数，L_{VAE}表示变分自动编码器损失函数；

[0139] 最后，利用得到的最终的损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络。

[0140] S105，利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0141] 本实施例中，所述基于卡尔曼滤波的位姿估计网络(KF-PoseNet)在设计时由于参考了卡尔曼滤波的思想，多次估计之间存在时序上的关联，因此，本发明中的KF-PoseNet可以更好地适应帧缺失的情况；

[0142] 本实施例中，在训练时，将一个批次的视频图像序列中的所有图像输入到位姿估计网络与深度估计网络中，对位姿估计网络与深度估计网络进行训练；进一步地，针对视觉里程计中存在的可能的帧缺失的情况，将一个批次的视频图像序列中的所有图像输入到深度估计网络中，并将该批次的视频图像序列中的一帧或多帧图像置零后再输入位姿估计网络，对位姿估计网络与深度估计网络再进行训练。例如，当N为5时，一个批次同时输入连续5帧图像到深度估计网络，并分别将每相邻两帧输入位姿估计网络；进一步地，针对视觉里程计中存在的可能的帧缺失的情况，从一次性输入的连续五帧的后3帧中随机将两帧图像置零，再输入位姿估计网络，进行训练，而深度估计网络的输入依然为完整的图像。

[0143] 在训练完成后，利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0144] 本发明提供的所述基于卡尔曼位姿估计网络的单目视觉里程计，能够较为有效地根据输入图像序列估计每一帧对应的相机位姿并适应帧缺失的情况。本发明适用于用于自监督单目视觉里程计。

[0145] 本发明实施例所述的基于卡尔曼位姿估计网络的单目视觉里程计方法，至少具有以下优点：

[0146] (1)针对现有的许多自监督方法没有考虑帧与帧之间的关联，帧间的信息没有被充分利用，导致训练出的网络难以估计出更为精确的位姿，也不能适应帧缺失的情况的问题，本实施例构建了基于卡尔曼滤波的位姿估计网络，并以此为基础，设计了针对帧缺失情况的训练策略，使得位姿估计网络可以利用帧间的信息估计当前的位姿，更加适应帧缺失的情况；

[0147] (2)针对场景中可能存在的运动物体与场景的欧氏变换不一致，不满足静态场景

的假设，难以用一个欧氏变换去描述场景的运动，导致位姿估计网络的估计结果出现偏差的问题，本实施例采用一种运动加权策略，同时在位姿估计网络与深度估计网络中引入变分自动编码器结构，使得位姿估计网络与深度估计网络可以在训练阶段更关注场景中的静止物体，提高网络泛化能力，提升网络性能。

[0148] 为了验证本发明实施例提供的基于卡尔曼位姿估计网络的单目视觉里程计方法的有效性，使用KITTI里程计数据集中提供的评估指标测试其性能：

[0149] (1)相对位移均方误差(Rel.trans.)：一个序列中全部长度为100、200、……、800米的子序列的平均位移RMSE(Root Mean Square Error)，以％度量，即每100米偏差的米数，数值越小越好。

[0150] (2)相对旋转均方误差(Rel.rot.)：一个序列中全部长度为100、200、……、800米的子序列的平均旋转RMSE，以deg/m度量，数值越小越好。

[0151] 本实施例中，应用了KITTI里程计数据集中00-07这八个序列作为训练集与验证集训练位姿估计网络与深度估计网络，并用09-10这两个序列来测试所述的用于自监督单目视觉里程计的基于卡尔曼滤波的位姿估计网络的性能。

[0152] KITTI里程计数据集是车载相机等设备采集的城市中公路环境的双目图像，雷达点以及实际轨迹。

[0153] 在实施过程中，构建深度估计网络和基于卡尔曼滤波的位姿估计网络；其中，位姿估计网络，用于输出输入的每对相邻帧图像之间的位姿变换，深度估计网络，用于输出输入帧的深度图像；根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像，计算视频图像序列基于运动加权的光度误差损失函数；在构建的位姿估计网络与深度估计网络中，引入变分自动编码器结构，计算变分自动编码器损失函数；基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络；利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿。

[0154] 在本实施例中，光度误差损失函数的超参数的参数$\alpha_0=0.85$，$\alpha_1=0.1$，$\alpha_2=0.05$，深度平滑损失函数的参数$\lambda=10^{-3}$，VAE损失函数参数$\lambda_1=\lambda_2=0.01$。网络的训练过程中，初始学习率为10^{-4}，并随着训练的进行逐渐减小，每经过一轮迭代，学习率变为上一轮的0.97倍，采用Adam优化器进行45次迭代，每轮迭代的批量大小为2，每批次包含连续N＝3帧图像。

[0155] 为了验证本发明所述方法的性能，本实施例中，选择了近几年基于深度学习的自监督的单目视觉里程计方法进行了对比，实验结果如表2所示。本实施例生成的轨迹如图4所示，其中，虚线轨迹为真实的轨迹，实线轨迹为本实施例中估计出的轨迹。

[0156] 由表2可以看出，相比于其他方法，由于对过去时刻提取出的信息的更好利用，对运动像素的加权，以及对VAE结构的应用，本实施例所述的方法取得了更好的性能。

[0157] 表2本实施例的方法与其他方法对比

[0158]

方法	09 Rel. trans. (%)	09 Rel. rot. (deg/m)	10 Rel. trans. (%)	10 Rel. rot. (deg/m)
SfMLearner	8.28	0.031	12.2	0.03
GeoNet	28.72	0.098	23.9	0.09
DepthVOFeat	11.93	0.039	12.45	0.035
Vid2depth	-	-	21.54	0.125
UnDeepVO	7.01	0.036	10.63	0.046
Wang et al.	9.88	0.034	12.24	0.052
Ranjan et al.	6.92	0.018	7.97	**0.031**
DeepMatchVO	9.91	0.038	12.18	0.059
PoseGraph	8.1	0.028	12.9	0.032
MonoDepth	11.47	0.032	**7.73**	0.034
SC-SfMLearner	11.2	0.034	10.1	0.05
Jau et al.	8.64	0.66	11.72	0.95
本实施例	**6.78**	**0.0267**	7.96	0.0448

[0159] 为了验证本实施例所述的方法各部分的意义，本实施例中还进行了消融实验。实验结果如表3所示，其中，第二行中的“without kalman struct”表示去除网络中的卡尔曼结构，此时位姿估计网络的解码器结构为四层卷积层，前三层卷积层的激活函数为ReLU，第四层输出经过全局平均池化得到6自由度位姿向量。第三至第五行的“without motion weighting”，“without VAE”，“without LTC”分别对应去除网络中的运动加权、VAE结构和长时一致性约束的实验结果。第六行和第七行的“#fc＝6”和“#fc＝2”分别表示位姿估计网络解码器部分采用不同层数的全连接层的实验结果。第一行“basic”表示不添加以上三个结构的实验结果。最后一行表示本文完整的方法的实验结果。

[0160] 从实验结果中可以看到，类似与卡尔曼的结构，使得网络在估计当前相邻帧时可以从之前的数据中得到参考，使得当前的估计结果更为精确；运动加权的引入，使得网络在训练的时候可以更关注环境中静止物体的像素，削弱了与相机欧氏变换不一致的物体的干扰；VAE结构的引入，使得网络的解码器对编码器的结果中的噪声更具有鲁棒性，提高了网络的泛化能力，使结果有了近一步的提高。最终本文完整的方法取得了更好的实验结果。我们的方法的性能随着各个部分的增加而逐渐上升，证明了我们的方法中各个部分的意义。

[0161] 表3消融实验结果

[0162]

方法	Seq. 09 Rel. trans. (%)	Seq. 09 Rel. rot. (deg/m)	Seq. 10 Rel. trans. (%)	Seq. 10 Rel. rot. (deg/m)
basic	13.87	0.0407	11.57	0.0497
without kalman struct	8.532	0.0343	8.260	**0.0367**
without motion weighting	10.358	0.0371	8.378	0.0438
without VAE	9.6810	0.0385	9.774	0.0470
without LTC	8.6230	0.0359	15.368	0.0620
#fc=6	11.376	0.0356	10.586	0.0492
#fc=2	7.5520	0.0289	12.291	0.0607
ours	**6.78**	**0.0267**	**7.96**	0.0448

[0163] 表4帧缺失的情况的实验结果

[0164]

方法	Seq. 09 Rel. trans. (%)	Seq. 09 Rel. rot. (deg/m)	Seq. 10 Rel. trans. (%)	Seq. 10 Rel. rot. (deg/m)
without frame missing training	48.621	0.1241	39.897	0.1858
without kalman struct	13.006	0.0441	13.083	0.0626
ours	**8.440**	**0.0332**	**10.230**	**0.0386**

[0165] 本实施例还对本发明中设计的针对帧缺失情况的训练策略进行了消融实验。在测试时，本实施例采取第50、150……帧将一帧图像置零，第100、200……帧将两帧置零的方式，对本发明在帧缺失的情况进行测试。测试结果如表4所示。其中，第一行“without frame missing training”表示本实施例中在训练时不采用针对帧缺失情况的训练方法进行训练的结果，第二行“without kalman struct”表示采用针对帧缺失情况的训练方法但不采用卡尔曼结构的实验结果，第三行为本实施例中采取针对帧缺失情况的训练方法训练的实验结果。从表4中可以看出，本实施例提出的方法可以很好地适应帧缺失的情况。

[0166] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

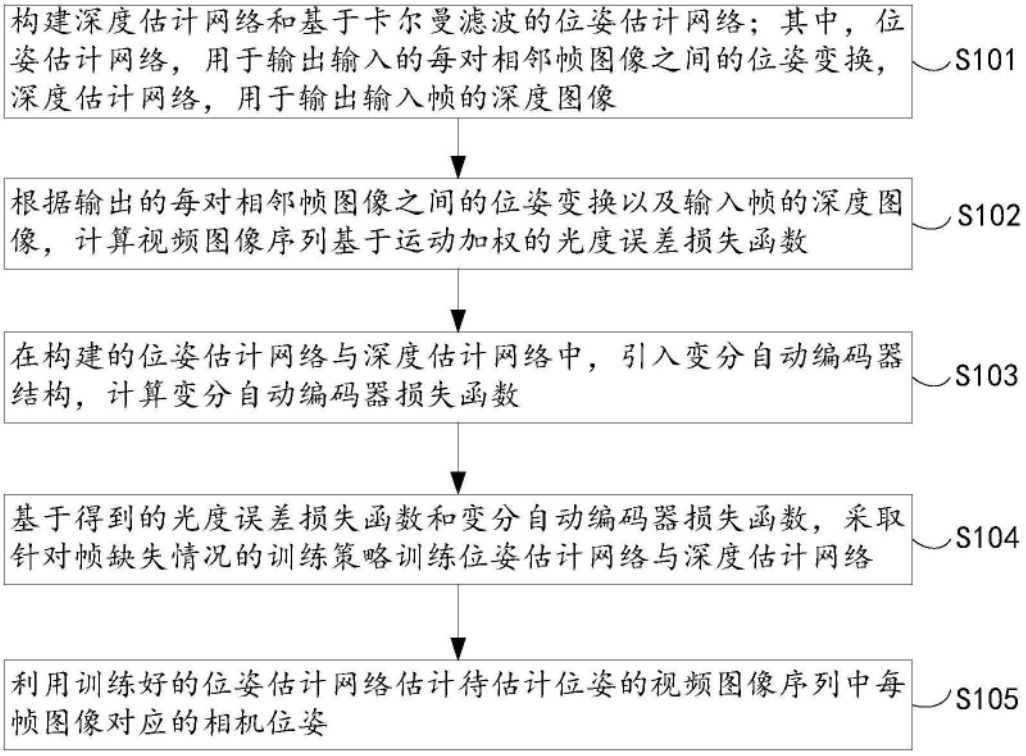

构建深度估计网络和基于卡尔曼滤波的位姿估计网络；其中，位姿估计网络，用于输出输入的每对相邻帧图像之间的位姿变换，深度估计网络，用于输出输入帧的深度图像
S101
根据输出的每对相邻帧图像之间的位姿变换以及输入帧的深度图像，计算视频图像序列基于运动加权的光度误差损失函数
S102
在构建的位姿估计网络与深度估计网络中，引入变分自动编码器结构，计算变分自动编码器损失函数
S103
基于得到的光度误差损失函数和变分自动编码器损失函数，采取针对帧缺失情况的训练策略训练位姿估计网络与深度估计网络
S104
利用训练好的位姿估计网络估计待估计位姿的视频图像序列中每帧图像对应的相机位姿
S105

图1

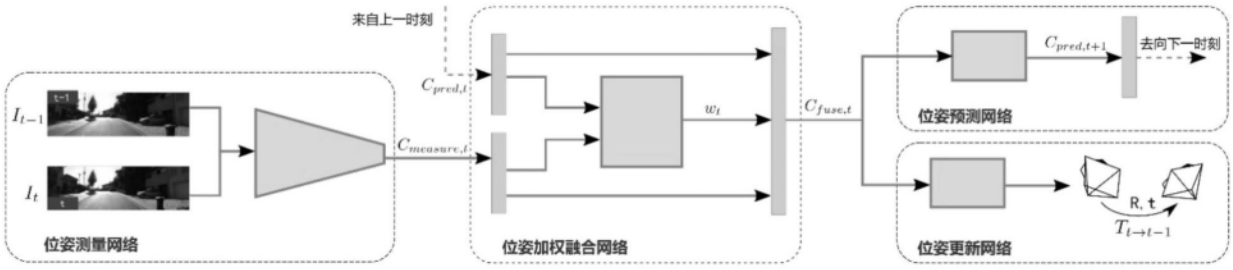

来自上一时刻
I_{t-1}
I_t
$C_{pred,t}$
$C_{measure,t}$
w_t
$C_{fuse,t}$
$C_{pred,t+1}$
去向下一时刻
位姿测量网络
位姿加权融合网络
位姿预测网络
位姿更新网络
R, t
$T_{t\to t-1}$

图2

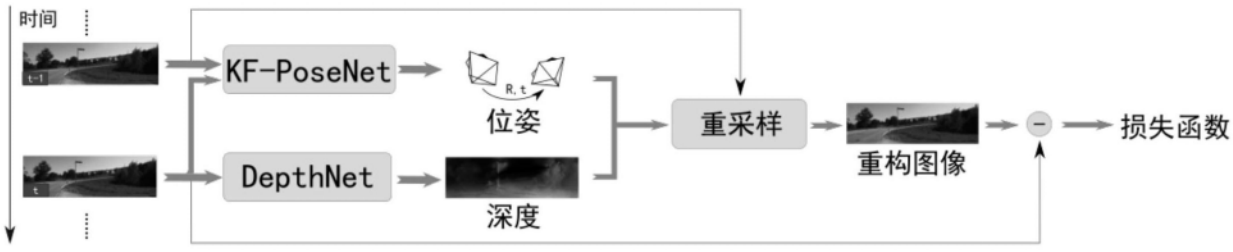

时间
KF-PoseNet
DepthNet
R, t
位姿
深度
重采样
重构图像
损失函数

图3

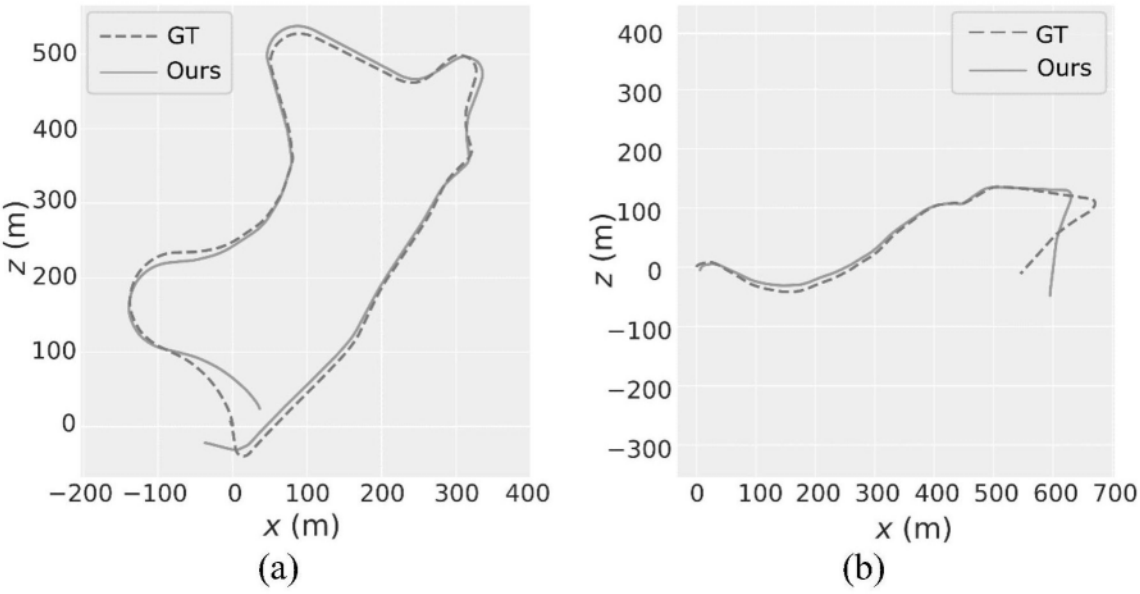


图4