(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 108230240 B
(45)授权公告日 2020.07.31

(21)申请号 201711493997.9

(22)申请日 2017.12.31

(65)同一申请的已公布的文献号
申请公布号 CN 108230240 A

(43)申请公布日 2018.06.29

(73)专利权人 厦门大学
地址 361005 福建省厦门市思明南路422号

(72)发明人 纪荣嵘 郭锋 黄剑波

(74)专利代理机构 厦门南强之路专利事务所（普通合伙）35200
代理人 马应森

(51)Int.Cl.
G06T 3/00(2006.01)
G06T 7/73(2017.01)
G06N 3/08(2006.01)
G06K 9/62(2006.01)

(56)对比文件
CN 104751184 A,2015.07.01
CN 105389550 A,2016.03.09
CN 106250931 A,2016.12.21
CN 103761526 A,2014.04.30
CN 105005789 A,2015.10.28
CN 106203354 A,2016.12.07
F.Walch,et al..“Image-based localization using LSTMs for structured feature correlation”.《2017 IEEE International Conference on Computer Vision》.2017,
Han Chen,et al..“Optimization Algorithm Toward Deep Features Based Camera Pose Estimation”.《International Conference on Images and Graphics》.2017,
任艺.“基于LDA主题模型的图像场景分类研究”.《中国优秀硕士学位论文全文数据库 信息科技辑》.2017,
Eric Brachmann,et al..“Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image”.《2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)》.2016,

审查员 易建琼

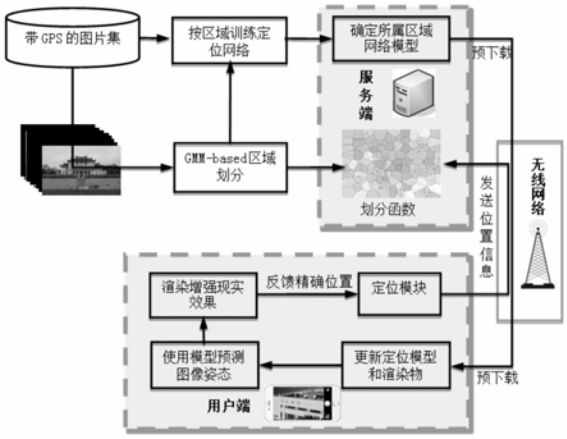
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于深度学习获取图像城市范围内位置及姿态的方法

(57)摘要

一种基于深度学习获取图像城市范围内位置及姿态的方法，涉及图像地理位置识别和增强现实领域。包括如下步骤：1)创建城市图片集；2)对城市图片集训练混合高斯模型，用训练出的混合高斯模型划分城市地理区域；3)训练联合学习图片姿态估计和场景识别神经网络；4)初始化，上传用户的GPS或者网络粗略位置信息；5)使用学习的划分函数对粗略的位置信息进行划分，下载对应网络模型和需要展示的渲染资料到用户端；6)采集用户输入相机视频流，运用下载的当前区域的网络模型预测当前时刻的三个层面的定位结果，若网络输出的预测结果置信度高于阈值，则使用预测的位置和姿态参数进行渲染资料的渲染。



CN 108230240 B

1.一种基于深度学习获取图像城市范围内位置及姿态的方法,其特征在于包括如下步骤:

1)创建城市图片集;

2)对城市图片集训练混合高斯模型,用训练出的混合高斯模型划分城市地理区域:初始化的城市图片集训练混合高斯模型,用训练出的混合高斯模型划分城市地理区域,初始化的图片数据集包含了M个地标区域$c_{1,2...M}$,位置标签x_j属于某个区域c_i,使用第j张图片属于第i个区域的后验概率来确定图片j属于哪个区域;具体步骤为:

(1)用贝叶斯公式计算后验概率:

$$p(y=i\mid x_j)=\frac{p\left(x_j\middle|y=i\right)P\left(y=i\right)}{p\left(x_j\right)}$$

条件概率$p(y=i\mid x_j)$表示x_j属于区域c_i的概率,$p(x_j\mid y=i)$服从归一化的高斯分布:

$$p(x_j\mid y=i)=\frac{\exp[-\frac{1}{2}(x_j-u_i)^T\sum_i^{-1}(x_j-u_i)]}{(2\pi)^{\frac{M}{2}}\|\Sigma_i\|^{\frac{1}{2}}}$$

其中x_j-u_i表示照片j与第i类区域中心之间的地理距离;

(2)由于各个成分的参数和每张图片的区域分配都是未知的,因此采用EM算法求解混合高斯模型,对区域划分概率密度函数进行参数估计,对数似然函数的计算如下:

$$Likelihood=\sum_{j=1}^{J}\log\sum_i^M\theta_i p\left(x_j\mid y=i\right)p(y=i)$$

θ_i是第i个高斯成分的系数,系统算法在EM过程逐步迭代逼近最大似然值;

(3)在第t次迭代估计一个GMM模型的参数λ_t:

$\lambda_t=\{\mu_1(t),...\mu_M(t),\Sigma_1(t),...\Sigma_M(t),p_1(t),...p_M(t)\}$

(4)设置λ_t对每个训练样本计算似然函数,随后用似然函数最大的分布更新参数λ_{t+1};

(5)重复计算步骤(3)和(4),直到似然函数的值收敛为止,算法得到对于样本x_j的最优区域指派$p(y=i\mid x_j,\lambda_t)$以及对应高斯成分的最优参数;

3)训练联合学习图片姿态估计和场景识别神经网络,具体方法为:在网络末端引出三个子网络,同时进行训练,第一个输出和第二个输出分别用于输入图片位置的回归和图片方向的回归,使用欧式损失来训练姿态估计,为每张图片计算与ground truth的位置损失$Loss_{loc}$和方向损失$Loss_{oren}$如下:

$$Loss_{loc}=\sum_i^n\|x_i-\tilde{x}_i\|_2 \qquad (1)$$

$$Loss_{oren}=\sum_i^n\|q_i-\tilde{q}_i\|_2 \qquad (2)$$

三维向量x表示图像相机在三维空间中位置XYZ,四元素向量q表示三维空间中的方向,带head的变量表示样本的ground truth;

第三个子网络输出一个离散的概率分布,其中p_0代表背景类的概率,用于输出分类的最后一层全连接层拥有C+1个神经元,使用Softmax计算对应于每个类别的输出概率p_c,分类

任务的Softmax损失公式如下：

$$loss_{cls} = -\sum_{i}^{n}\sum_{c}^{C}\tilde{p}_i^c log\left(p_i^c\right) \quad (3)$$

$$p_i^c = \frac{\exp\left(p_i^c\right)}{\Sigma_c \exp\left(p_i^c\right)} \quad (4)$$

其中，p_i^c表示样本属于类别的概率，若样本属于类别，则标注$\tilde{p}_i^c=1$，否则等于$\tilde{p}_i^c=0$，最后计算出3个单独损失的加权求和来计算整个模型的总损失：

$$\mathcal{L}_{total} = \sum_{t=1}^{t=3}\lambda_t loss_t \quad (5)$$

计算公式(5)中的$loss_t$代表第t个损失函数，λ_t表示它们的权重系数，权重λ_t由各个任务在整个目标函数中的重要程度决定；

4)初始化，上传用户的GPS或者网络粗略位置信息；

5)使用学习的划分函数对粗略的位置信息进行划分，下载对应网络模型和需要展示的渲染资料到用户端；

6)采集用户输入相机视频流，运用下载的当前区域的网络模型预测当前时刻的三个层面的定位结果，若网络输出的预测结果置信度高于阈值，则使用预测的位置和姿态参数进行渲染资料的渲染。

2.如权利要求1所述一种基于深度学习获取图像城市范围内位置及姿态的方法，其特征在于在步骤1)中，所述创建城市图片集的具体方法为：使用网络爬虫至图片分享网站下载城市中各个地方的景观图片，组成城市级别的图片数据库；假设初始化图片数据集包含了M个地标区域$c_{1,2...M}$，位置标签x_j属于某个区域c_i。

3.如权利要求1所述一种基于深度学习获取图像城市范围内位置及姿态的方法，其特征在于在步骤3)中，所述位置包括区域经纬度范围(Lat,Lng)、所属建筑场景的Class ID、在建筑场景坐标系中的相对坐标(X,Y,Z)、相对参考视点的方向向量四元素(q,x,y,z)。

4.如权利要求1所述一种基于深度学习获取图像城市范围内位置及姿态的方法，其特征在于在步骤3)中，所述权重λ_t设为：$\lambda_{loc}=1$，$\lambda_{oren}=250$，$\lambda_{cls}=0.5$。

一种基于深度学习获取图像城市范围内位置及姿态的方法

技术领域

[0001] 本发明涉及图像地理位置识别和增强现实领域，尤其是涉及一种基于深度学习获取图像城市范围内位置及姿态的方法。

背景技术

[0002] 随着移动互联网和智能设备的爆发式发展，拍摄和分享照片已经成为人们日常的一部分。如何从照片中推断出照片被拍摄的位置和拍摄的视角成为一项很有意义的问题。从照片中推断出拍摄位置和视角的问题在立体视觉(Multi-View Stereo)中也称为相机的姿态估计问题，是计算机视觉和机器人领域中的一个基本问题，拥有广泛的应用，比如在增强现实(Augmented Reality，简称AR)，同时定位与地图构建(Simultaneous Localization and Mapping，简称SLAM)，以及基于图像的地理位置识别(Image-based Location Recognition)通过把虚拟的3D图像或标注以接近真实的观测视角重叠在相机视频流上，以加强用户对现实世界的理解，增强现实已经被广泛应用于工业领域和消费领域，包含教育、医疗、娱乐、设计、军事等等。增强现实模块需要实时地估计图像的相机姿态，以提供相机设备在3D空间中对于位置和角度的6个自由度的参数，这一过程也称为图像重定位，或者“外参数标定”(Extrinsic calibration)。姿态估计的速度和准确度直接影响了增强现实的渲染，进而决定了用户体验的优劣。

[0003] 虽然Weyand T([1]Weyand T,Leibe B.Discovering favorite views of popular places with iconoid shift[C]//International Conference on Computer Vision.IEEE Computer Society,2011:1132-1139)，Li X([2]Herranz L,Jiang S,Li X.Scene Recognition with CNNs:Objects,Scales and Dataset Bias[C]//Computer Vision and Pattern Recognition.IEEE,2016:571-579)和Larson M在基于图像场景识别和地点识别中做了许多优秀的工作，但是他们大都只单纯的进行地名识别或经纬度坐标估计。Shotton等([3]Shotton J,Glocker B,Zach C,et al.Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images[C]//IEEE Conference on Computer Vision and Pattern Recognition.IEEE Computer Society,2013:2930-2937)则以RGB-D图像作为输入，用深度图像创建场景坐标的标注，把每个像素从摄像机坐标系映射到全局的场景坐标系中。然后像素和预先标注好的场景模型之间的映射关系训练一个回归森林。场景坐标回归森林本质上是在学习映射函数。然而，该算法的输入是RGB-D图像，RGB-D图像仅适用于室内场景。在测试阶段，为了进行图像定位，首先将查询图片输入到训练好的森林进行回归计算，然后使用基于RANSAC的姿态验证来确定一个一致的相机姿态结果。尽管坐标回归森林的准确度非常高，但它的缺点主要是需要RGB-D图像作为输入，在实际使用中，RGB-D图像只适用于室内场景，并且RANSAC计算过程非常耗时。

发明内容

[0004] 本发明的目的在于提供一种基于深度学习获取图像城市范围内位置及姿态的方

法。

[0005] 本发明包括如下步骤：

[0006] 1)创建城市图片集；

[0007] 2)对城市图片集训练混合高斯模型，用训练出的混合高斯模型划分城市地理区域；

[0008] 3)训练联合学习图片姿态估计和场景识别神经网络；

[0009] 4)初始化，上传用户的GPS或者网络粗略位置信息；

[0010] 5)使用学习的划分函数对粗略的位置信息进行划分，下载对应网络模型和需要展示的渲染资料到用户端；

[0011] 6)采集用户输入相机视频流，运用下载的当前区域的网络模型预测当前时刻的三个层面的定位结果，若网络输出的预测结果置信度高于阈值，则使用预测的位置和姿态参数进行渲染资料的渲染。

[0012] 在步骤1)中，所述创建城市图片集的具体方法可为：使用网络爬虫至图片分享网站下载城市中各个地方的景观图片，组成城市级别的图片数据库；假设初始化图片数据集包含了M个地标区域$c_{1,2...M}$，位置标签x_j属于某个区域c_i。

[0013] 在步骤2)中，所述对城市图片集训练混合高斯模型，用训练出的混合高斯模型划分城市地理区域的具体方法可为：初始化的城市图片集训练混合高斯模型，用训练出的混合高斯模型划分城市地理区域，初始化的图片数据集包含了M个地标区域$c_{1,2...M}$，位置标签x_j属于某个区域c_i，使用第j张图片属于第i个区域的后验概率来确定图片j属于哪个区域；

[0014] (1)用贝叶斯公式计算后验概率：

[0015] $$p(y=i\mid x_j)=\frac{p\left(x_j\middle|y=i\right)P\left(y=i\right)}{p\left(x_j\right)}$$

[0016] 条件概率$p(y=i\mid x_j)$表示x_j属于区域c_i的概率，$p(x_j\mid y=i)$服从归一化的高斯分布：

[0017] $$p(x_j\mid y=i)=\frac{\exp[-\frac{1}{2}(x_j-u_i)^T\sum_i^{-1}(x_j-u_i)]}{(2\pi)^{\frac{M}{2}}\|\Sigma_i\|^{\frac{1}{2}}}$$

[0018] 其中x_j-u_i表示照片j与第i类区域中心之间的地理距离；

[0019] (2)由于各个成分的参数和每张图片的区域分配都是未知的，因此采用EM算法求解混合高斯模型，对区域划分概率密度函数进行参数估计，对数似然函数的计算如下：

[0020] $$Likelihood=\sum_{j=1}^{J}\log\sum_i^M\theta_i p\left(x_j\mid y=i\right)p(y=i)$$

[0021] θ_i是第i个高斯成分的系数，系统算法在EM过程逐步迭代逼近最大似然值；

[0022] (3)在第t次迭代估计一个GMM模型的参数λ_t：

[0023] $\lambda_t=\{\mu_1(t),...\mu_M(t),\Sigma_1(t),...\Sigma_M(t),p_1(t),...p_M(t)\}$

[0024] (4)设置λ_t对每个训练样本计算似然函数，随后用似然函数最大的分布更新参数λ_{t+1}；

[0025] (5)重复计算步骤(3)和(4)，直到似然函数的值收敛为止，算法得到对于样本x_j的

最优区域指派$p(y=i|x_j,\lambda_t)$以及对应高斯成分的最优参数。

[0026] 在步骤3)中，所述训练联合学习图片姿态估计和场景识别神经网络的具体方法可为：在网络末端引出三个子网络，同时进行训练，第一个输出和第二个输出分别用于输入图片位置的回归和图片方向的回归，使用欧式损失来训练姿态估计，为每张图片计算与ground truth的位置损失$Loss_{loc}$和方向损失$Loss_{oren}$如下：

[0027] $$Loss_{loc}=\sum_{i}^{n}\|x_i-\tilde{x}_i\|_2 \quad (6)$$

[0028] $$Loss_{oren}=\sum_{i}^{n}\|q_i-\tilde{q}_i\|_2 \quad (7)$$

[0029] 三维向量x表示图像相机在三维空间中位置XYZ，四元素向量q表示三维空间中的方向，带head的变量表示样本的ground truth；

[0030] 第三个子网络输出一个离散的概率分布，其中p_0代表背景类的概率，用于输出分类的最后一层全连接层拥有C+1个神经元，使用Softmax计算对应于每个类别的输出概率p_c，分类任务的Softmax损失公式如下：

[0031] $$loss_{cls}=-\sum_{i}^{n}\sum_{c}^{C}\tilde{p}_i^c log\left(p_i^c\right) \quad (8)$$

[0032] $$p_i^c=\frac{\exp\left(p_i^c\right)}{\Sigma_c\exp\left(p_i^c\right)} \quad (9)$$

[0033] 其中，p_i^c表示样本属于类别的概率，若样本属于类别，则标注$\tilde{p}_i^c=1$，否则等于$\tilde{p}_i^c=0$，最后计算出3个单独损失的加权求和来计算整个模型的总损失：

[0034] $$\mathcal{L}_{total}=\sum_{t=1}^{t=3}\lambda_t loss_t \quad (10)$$

[0035] 计算公式(5)中的$loss_t$代表第t个损失函数，λ_t表示它们的权重系数，权重λ_t由各个任务在整个中的重要程度决定。

[0036] 所述位置可包括区域经纬度范围(Lat,Lng)、所属建筑场景的(Class ID)、在建筑场景坐标系中的相对坐标(X,Y,Z)、相对参考视点的方向向量四元素(q,x,y,z)等。

[0037] 所述权重λ_t可设置为：$\lambda_{loc}=1$，$\lambda_{oren}=250$，$\lambda_{cls}=0.5$。

[0038] 本发明解决如下应用场景：在一个城市中的用户用手机拍摄室外照片，应用要求定位这张照片，准确计算出拍摄地点、姿态，最后基于这些定位结果使用增强现实技术为用户渲染特定信息。因为在更大的场景中，获取准确的姿态标注难度增加，所以基于深度特征的图像重定位方法通常只能用于中小场景。当放大到城市级别的时候，训练集需要包含整个城市的所有表观，更合理的做法是仅对城市中一些热门的地区、地标进行提供定位服务。

[0039] 与现有技术相比，本发明的优点和积极效果是把传统二维方法的图像地理位置定位拓展到三维空间中，达到丰富图片拍摄位置信息的目的。通过机器学习算法来学习集合图片位置和图片视觉内容对城市地理地貌和建筑风格进行聚类、划分，达到“由粗到细、由大到小”的定位目的，从而解决了复杂的城市中图像的定位问题。首先描述图像在现实世界

中的位置包含了多层语义的关系，快速且精准的获取图像位置信息是LBS等应用的技术难点。本发明的技术提供多层次语义的精确地理位置描述可以拉近用户与物理世界的距离，降低用户认知物理空间的成本，有望为机器人、无人机和智能设备的自我定位问题找到新的解决方法。

附图说明

[0040] 图1为本发明的系统框架图。

[0041] 图2为本发明的渲染模块流程图。

[0042] 图3为本发明的联合相机姿态估计和场景识别神经网络框架。

具体实施方式

[0043] 下面结合实施例和附图进一步说明本发明。

[0044] 一、发明的整体流程设计

[0045] 本发明在PC端设计了基于深度学习获取图像城市范围内位置及姿态的实现系统，框架图如图1所示。整个发明的系统分为在线部分和在线部分。离线部分主要在服务器端，训练区域划分学习器把整个城市划分成一个个子区域，之后对每个子区域采用迁移学习的方法训练提出的姿态回归和场景分类网络。对在线部分主要在移动客户端，用户到达某个区域后给服务器发送GPS或者手机基站的地理位置，服务器根据区域划分学习器的划分函数确定用户所属区域(场景)，用户下载所属区域的网络模型和需要展示的信息。运行网络模型，输入照片并输出准确的位置坐标、姿态参数以及场景类别，根据输出的姿态渲染需要展示的信息。最后用户端返回估计结果，服务器记录用户准确的定位结果。

[0046] 在实际使用阶段，用户只需下载所属建筑区域的网络模型。化大为小，按需更新的好处是，用户不需要下载整个城市的模型节省了时间和通信成本。同时，根据聚类结果划分小区域后，训练出的网络模型对该区域的姿态估计误差更低，因为对模型泛化性能要求降低了。不同于地标识别系统，我们的系统得到的是一个由粗到细、包含4个层次的位置描述。

[0047] 所述位置描述包含区域经纬度范围(Lat,Lng)。

[0048] 所述位置描述包含:所属建筑场景的(Class ID)。

[0049] 所述位置描述包含在建筑场景坐标系中的相对坐标(X,Y,Z)。

[0050] 所述位置描述包含相对参考视点的方向向量四元素(q,x,y,z)。

[0051] 二、深度卷积神经网络预测图像位置和姿态

[0052] 本发明在使用阶段把网络中的输入层和loss都去掉。输入一张图crop成224×224图像后，输入重定位网络，将得到图像的三维空间位置、代表方向的四元素向量、图像最可能的场景类别及置信度。

[0053] 本发明首先从摄像头获取视频流，缩放读入帧的最短边位256，然后在图像中间裁剪224×224的图像，变换为BGR颜色通道。之后调用CAFFE开源库加载神经网络模型的参数。输入裁剪图片，经过一系列卷积运算得到视频帧的特征图。对特征图进行连续卷积和下采样，最后连接多个特征图，输出到图像的三维空间位置XYZ、代表方向的四元素向量Q、图像最可能的场景类别L及预测结果的置信度C。

[0054] 三、对定位结果的渲染方法

[0055] 本发明在渲染定位结果时，采用了两个线程并行处理。

[0056] 1.姿态估计线程不断的接收新的相机帧，并运行本设计的深度学习算法预测输出场景类别和置信度以及对应相机姿态。置信度如果超过0.7，选择保留当前帧的姿态估计结果。把姿态结果输入卡尔曼滤波器，平滑姿态估计的抖动。把平滑之后的6个自由度姿态结果转换成相机视角矩阵ModelView矩阵。传递到OpenGL渲染线程。

[0057] 2.OpenGL线程用于绘制用户界面和渲染增强现实效果。OpenGL线程根据当前确定的场景类别确定要在什么位置放什么叠加物，并送到OpenGL管线中。OpenGL线程会一直读取视角矩阵变量，如果姿态估计线程传递了新的摄像机视角矩阵，对应的渲染视角会发生变化，从而达到增强现实效果。

[0058] 在Ubuntu14.04对本发明的算法进行实现与测试，调试使用单目摄像头来进行。使用OpenCV打开摄像头读取视频流，对视频帧进行操作。使用CAFFE库运行姿态估计线程。使用OpenGL进行渲染，使用GLUT绘制软件的UI界面，并利用JNI技术在Android 6.0操作系统上进行移植工作，系统中获取视频流、视频图像处理、界面以及OpenGL渲染模块，从而实现整个发明。

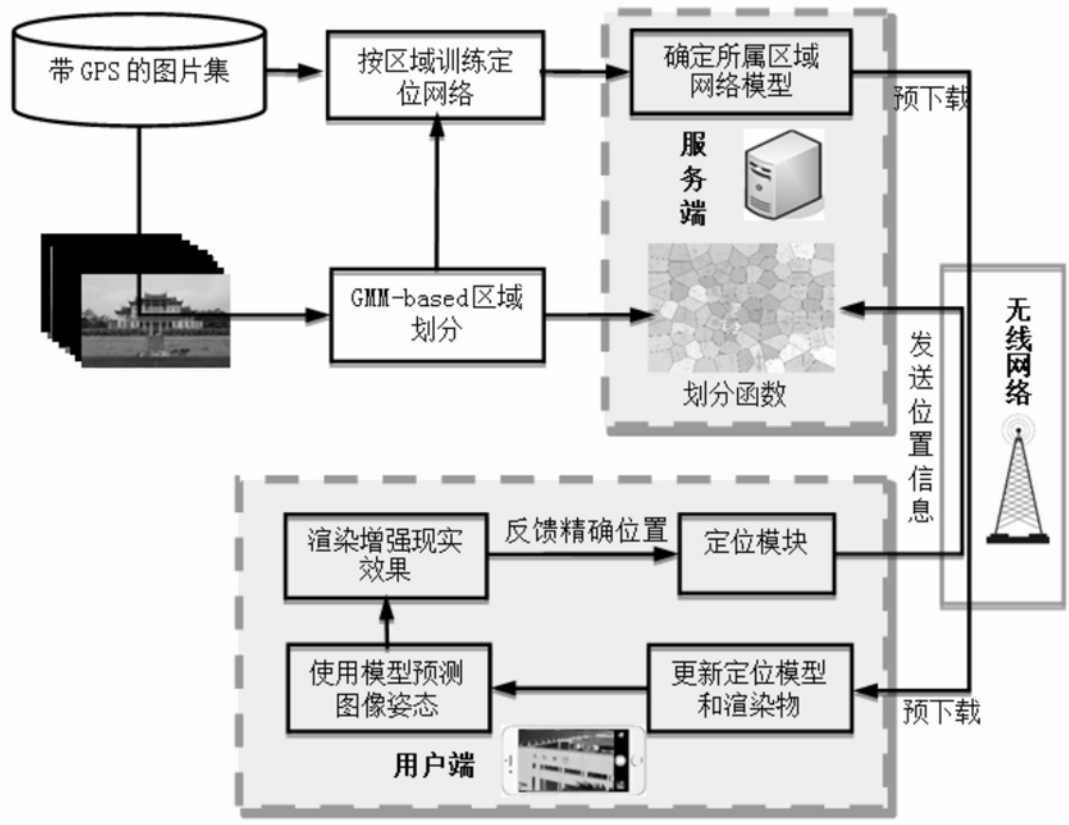

带GPS的图片集
按区域训练定位网络
确定所属区域网络模型
预下载
服务端
GMM-based区域划分
划分函数
发送位置信息
无线网络
渲染增强现实效果
反馈精确位置
定位模块
使用模型预测图像姿态
更新定位模型和渲染物
预下载
用户端

图1

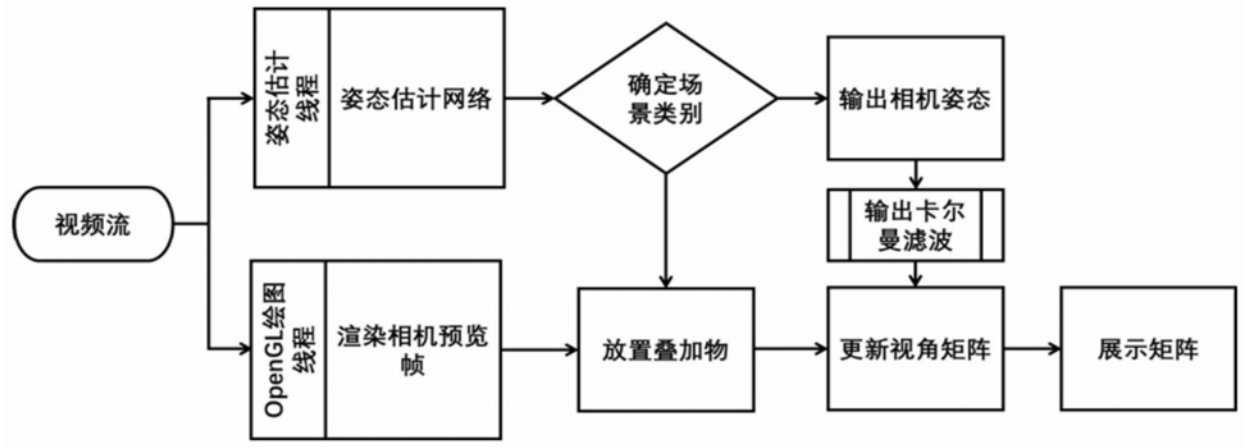

视频流
姿态估计线程
姿态估计网络
确定场景类别
输出相机姿态
输出卡尔曼滤波
OpenGL绘图线程
渲染相机预览帧
放置叠加物
更新视角矩阵
展示矩阵

图2

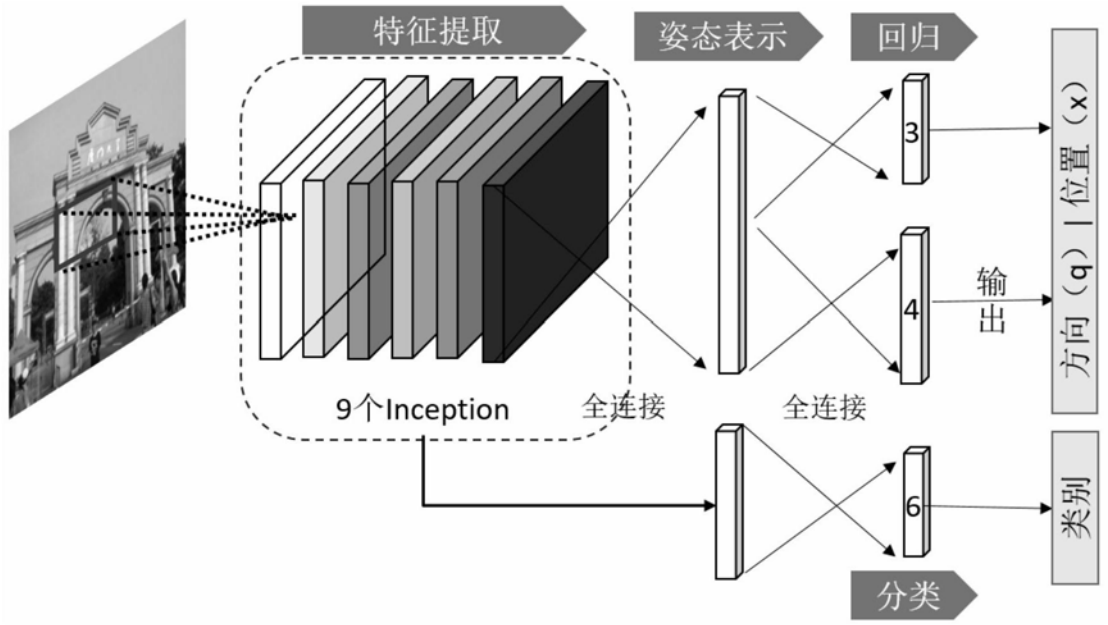


图3