计算机科学
COMPUTER SCIENCE

基于双目估计的动态场景三维感知技术研究与实现

何维龙, 苏玲莉, 郭丙轩, 李茂森, 郝岩

引用本文

何维龙, 苏玲莉, 郭丙轩, 李茂森, 郝岩. 基于双目估计的动态场景三维感知技术研究与实现[J]. 计算机科学, 2024, 51(11A): 240300045-8.

HE Weilong, SU Lingli, GUO Bingxuan, LI Maosen, HAO Yan. Research and Implementation of Dynamic Scene 3D Perception Technology Based on BinocularEstimation [J]. Computer Science, 2024, 51(11A): 240300045-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300045

基于双目估计的动态场景三维感知技术研究与实现

何维龙[1]　苏玲莉[1]　郭丙轩[2]　李茂森[3]　郝　岩[1]
1 酒泉职业技术学院　甘肃 酒泉 735000
2 武汉大学测绘遥感信息工程国家重点实验室　武汉 430072
3 核工业航测遥感中心　河北 保定 071799
(1205488897@qq.com)

摘　要　双目立体视觉技术在计算机视觉领域研究中一直具有重要意义。不同于单目或多目技术,双目立体视觉在能够准确获取图像深度的同时,也兼具了低成本、高泛用性、使用简便等优势。基于双目视觉的三维感知技术能够极大提升计算机对现实世界的理解和交互能力,进一步增强计算机视觉技术在复杂、多变的场景中的适应能力,在自动驾驶、机器人导航、工业检测、航天等领域发挥着重要作用。文中重点研究动态场景中的三维重建与目标感知技术,在大多数情况中,视野中的动态目标实际上是需要重点关注的目标,而静态目标,特别是在场景中绝大多数时候都占据主要空间的背景以及静态物体往往是可以被忽略掉的,但是在实际计算时确占用了大量资源。在场景中不受关注的目标上花费过多计算资源,显然是无意义且非常低效的。针对这个问题,本文在深入研究了目前主流的双目立体匹配方法、图像分割等方法的基础上,提出了一种基于双目估计的动态场景三维感知技术。主要的创新点和研究成果包括:针对传统双目立体匹配算法逐像素计算聚合低价效率低下的问题,提出了一种基于二维场景实例分割的双目立体匹配方法,使用 mask 分割后的目标图像进行立体匹配,这样不仅提升了匹配性能,同时也降低了动态目标的匹配难度。针对分割精确不足的问题,引入基于 RGB 图像的 mask 边缘滤波优化方法,在提升效率的同时提升视场点云重建精度。其次,基于双目估计深度学习网络进行实时目标点云生产,并提出基于 GPU 加速的邻近帧点云的实时动态目标感知算法。最后提出二三维一体的动态目标实时感知技术,在对目标场景实现实时三维重建的同时,快速识别检测环境中的动态目标物体。

关键词:双目视觉;立体匹配;图像分割;三维重建;深度学习;GPU 并行计算
中图分类号　P231

Research and Implementation of Dynamic Scene 3D Perception Technology Based on Binocular Estimation

HE Weilong[1],SU Lingli[1],GUO Bingxuan[2],LI Maosen[3] and HAO Yan[1]
1 Jiuquan Vocational and Technical College,Jiuquan,Gansu 735000,China
2 State Key Laboratory of Information Engineeringin Surveying Mapping and Remote Sensing,Wuhan University,Wuhan 430072,China
3 Nuclear Industry Aerial Surveying and Remote Sensing Center,Baoding,Hebei 071799,China

Abstract　Binocular stereo vision technology has always been of great significance in the field of computer vision research. Unlike monocular or multicular technology,binocular stereo vision has the advantages of low cost,high versatility,simple use and so on while it can accurately obtain the image depth. The three-dimensional perception technology based on binocular vision can greatly improve the computer's understanding and interaction ability to the real world,further enhance the adaptability of computer vision technology in complex and changeable scenes,and play an important role in the fields of automatic driving,robot navigation,industrial inspection,aerospace,etc. This paper focuses on 3D reconstruction and object perception technology in dynamic scenes. In most cases,dynamic objects in the field of vision usually need to be focused on,while static objects,especially the background and static objects in the scene that occupy the main space in most cases,can be ignored,but they do occupy a lot of resources in the actual calculation,It is obviously meaningless and inefficient to spend too much computing resources on targets that are not concerned in the scene. In order to solve this problem,based on the in-depth study of the current mainstream binocular stereo matching methods,image segmentation and other methods,this paper proposes a dynamic scene 3D perception technology based on binocular estimation. The main innovations and research achievements include:Aiming at the low cost and efficiency of the traditional binocular stereo matching algorithm in pixel by pixel computing aggregation,a binocular stereo matching method based on two-dimensional scene instance segmentation is proposed,and the target image after mask segmentation is used for stereo matc-

基金项目:国家重点研发计划(2019YFE0108300);国家自然科学基金(62001058);2023 年甘肃省高等学校创新基金项目(2023B-449);2023 年酒泉市科技支撑项目(2060499);校级科研项目(2022XJYXM06)
This work was supported by the National Key Research and Development Program of China(2019YFE0108300),National Natural Science Foundation of China(62001058),2023 Gansu Province Higher Education Innovation Fund Project (2023B-449),2023 Jiuquan City Science and Technology Support Project(2060499) and School level Scientific Research Project(2022XJYXM06).
通信作者:苏玲莉(479679364@qq.com)

hing, which not only improves the matching performance but also reduces the difficulty of dynamic target matching. At the same time, in order to solve the problem of insufficient segmentation accuracy, the mask edge filtering optimization method based on rgb image is introduced to improve the efficiency and the reconstruction accuracy of the field of view point cloud. Secondly, real-time target point cloud production is carried out based on binocular estimation depth learning network, and a real-time dynamic target perception algorithm based on GPU accelerated neighboring frame point cloud is proposed. At last, a two-dimensional and three-dimensional dynamic object real-time perception technology is proposed, which can quickly recognize the dynamic object in the detection environment while realizing real-time three-dimensional reconstruction of the target scene.

Keywords Binocular vision, Stereo matching, Image segmentation, 3D reconstruction, Depth learning, GPU parallel computing

在信息技术高速发展的背景下，双目立体视觉系统作为计算机视觉领域内极为重要的一部分，在自动驾驶、虚拟现实、载人航天、医学诊断、材料加工等各个领域具有极其广泛的应用[1-4]。而随着各领域对立体视觉的需求不断增大，如何进一步提升算法的精确度，减少计算的复杂程度，提高算法鲁棒性，是相关领域从业人员必须要面对、深入思考和探索的问题。

20 世纪 70 年代，Marr 教授[5]首次提出了一整套完善的计算机视觉领域基本理论框架，确定了由二维图像作为基础构建三维场景的理论依托和整体流程，为后续更加系统性地开展科研工作奠定了坚实的理论基础。随后，Barnard[6]对双目立体视觉的一整套理论框架进行了完善，他将整个双目立体视觉的完整过程分为了 6 个阶段，并进行详细叙述。后来随着研究工作的进一步发展，技术流程更加规范，双目立体视觉的总体步骤可以归纳为以下 4 个环节：图像输入、立体矫正、视差计算（立体匹配）以及三维重建，到目前为止这也是使用双目立体视觉法进行三维重建的标准流程[7]。2006 年，Yoon 等[8]提出一种自适应支持权值方法，借鉴了双边滤波的思想，根据色彩关系和空间距离调整匹配窗口内像素点的权重，这种方法有不错的匹配精度，但算法的计算量更大。同年，Gerris 等[9]首次在立体匹配算法中引入视觉图像分割，从而改善算法因图像存在较多重复纹理区域而出现的精度下降问题。随着近几年双目立体匹配在深度学习领域的快速发展，上述传统的视差计算方法已经逐渐被基于卷积神经网络的方法所取代。Laskowski[10]于 2013 年最早提出了基于双目立体视觉的神经网络算法，用于预测视差。2015 年，Zbontar 等[11]提出 MC-CNN 网络，他使用了一个孪生网络来提取输入图像的特征从而判断两幅图像的相似性，并将计算得到的相似性度量作为匹配代价，结合传统的视差计算和优化方案得到最终的视差图。该网络计算得到的视差结果精度极高，但是因为网络本身存在缺陷，导致处理速度非常慢，很难用于实际应用中。2016 年，Mayer 等[12]提出了 DispNet，这是第一个使用端到端网络实现双目立体匹配的方法，其通过构建编码器和解码器，提取特征和学习匹配代价，完成视差计算任务。后续的许多研究在 DispNet 的基础之上做出了改进，Pang 等[13]设计了一种二阶立体匹配网络，包括一个用于获取初始视差的第一阶网络，和一个用于计算不同尺度下初始视差的残差修正值的第二阶段网络，最后整合两个网络的输出得到最终视差值。2017 年，Kendall 等[14]提出了一个使用 argmin 函数实现视差计算方法的深度神经网络 GC-Net，并且第一次将目标物体的几何特性转化为一个具有深度信息的匹配代价。Chang 等[15]提出 PSM-Net，引入金字塔池化模块（Spatial Pyramid Pooling，SPP）来聚合多尺度的信息，并提出堆叠漏斗网络来进行特征提取与代价体构建。Tonioni 等在 2020 年提出的 RTS2Net[16]是第一个实现实时语义分割的双目立体匹配网络模型。

本文主要研究基于双目视觉估计的三维点云重建与动态目标感知，采取的技术路线是图像实例分割结合双目立体几何的方式。本文在基于双目估计的动态场景三维感知关键技术研究中的主要工作与创新点总结如下：

（1）使用深度学习网络进行二维场景的实时分割与提取，得到剔除指定背景的目标 mask，提高双目立体匹配效率；

（2）提出基于图像 RGB 引导的 mask 滤波与再优化，提升视场点云重建精度；

（3）基于双目立体几何深度学习网络进行实时目标点云生产，并提出基于 GPU 加速的邻近帧点云的实时动态目标感知算法。

1 基于实时图像分割的双目立体匹配

目前的主流双目立体匹配方法大多是基于图像中的像素点，而基于图像分割的双目立体匹配方法是以分割后的图像块作为基准来进行立体匹配，这样不仅提高了匹配性能，同时也降低了匹配难度。本文沿袭基于图割的匹配方法，同时为了解决图像分割算法可能出现的对于目标边缘分割效果不够精细的问题，提出引入 Canny 算子的 mask 滤波优化方法，对基于网络输出的视差图生成粗点云进行滤波优化，最终得到精细点云。

1.1 图像分割

传统的图像分割技术是早期比较流行的分割手段，大多作为辅助手段，用于在视觉领域的图像预处理阶段提高对图像进行分析的效率，主要包括基于阈值[17-19]、边缘[20]、区域[21]等的图像分割法。2000 年后，基于超像素[22-24]的图像分割方法开始出现。

但传统的分割方法仅仅基于图像的表层信息来进行分割，虽然具有不错的分割精度，但是实际的任务场景往往需要根据目标的语义信息来进行划分，传统方法对于这一类要求更高的分割任务是无法胜任的。随着卷积神经网络的加入，图像分割不再局限于像素点的特征信息，而是结合具体的图像语义，实现更为复杂、精准的像素级分割，即语义分割（SemanticSegmentation）。2015 年出现的全卷积神经网络（FCN[25]）是第一个实现了图像语义分割的端到端深度神经网络，它的出现让语义分割达到了像素级分割的水平，后续的一系列工作都是以 FCN 为基本框架进行的。金字塔场景解析网络（Pyramid Scene Parseing Network，PSPNet[26]）针对 FCN 在场景解析上存在分类错误、小目标丢失等问题，做了一系列改进。DeepLab 系列是谷歌团队在 2014 年至 2018 年提出的一系列语义分割模型[26]，其核心是通过在卷积核中加入值为 0 的“空洞”，在不额外增加模型参数和计算量的前提下，提升卷积核的感受野，称为空洞卷积（atrous 卷积）。

实例分割（Instance Segmentation）工作是在语义分割划

分类型的基础之上进一步细分。实例分割不仅能够区分场景中的对象类型，还能够划分出同一类型的每一个对象，因此实例分割具有与目标检测类似的特性，区别在于实例分割能够同时输出目标的类型与 Mask。Mask-RCNN[26] 可以说是最具代表性的实例分割算法，其沿用了 FasterR-CNN[27] 的思想，在基础特征网络之后又加入了全连接的分割网络，由原来的两个任务（分类＋回归）变为了 3 个任务（分类＋回归＋分割），同时进行高质量的目标检测与分割（见图 1）。

图 1　Mask-RCNN 实例分割效果

Fig. 1　Mask-RCNN instance segmentation effect

Mask R-CNN 与 Faster R-CNN 一样是一个两阶段的框架，第一阶段划分区域，第二阶段预测边界框和分割 Mask。

Mask R-CNN 在第一阶段具有和 Faster R-CNN 相同的第一层，即区域建议网络（Region Proposal Networks，RPN）。RPN 是一个轻量级的网络，通过滑动窗口来扫描图像，并提取出可能存在目标物体的感兴趣区域 ROI。经过 RPN 网络提取出的区域可能具有不同的尺寸，但是在输入到分类器时必须具有相同的输入尺寸，因此还需要通过 ROI 池化来对图像进行裁剪，将其重新调整为统一的尺寸。最终会将 ROI 划分为大小为 $m\times m$ 的一个个子区域。

第二阶段除了原有的预测种类和边界框回归任务外，为了提取目标的 Mask，添加了一个全卷积网络的分支与目标检测任务并行（实际上就是一个并行的 FCN 层），为每一个 ROI 预测二分类的掩码（binary mask），来表示某一像素是否属于目标的一部分，当像素是属于目标上的某一像素点时标识为 1，其他位置标识为 0。这里进行的分割实际上是语义分割而非实例分割，但是由于每一个 ROI 都仅对应于一个物体，也就意味着完成了语义分割就实现了实例分割的任务。换言之，MaskR-CNN 的整体实现框架是先分类后分割。

在 Mask R-CNN 网络的训练阶段，作者定义了一种多任务损失约束 L，损失函数由 3 部分组成：分类损失函数、边界框损失函数以及掩码 Mask 损失函数之和。L 的表达式如下：

$$L=L_{\text{cls}}+L_{\text{box}}+L_{\text{mask}} \tag{1}$$

其中，L_{cls} 和 L_{box} 的定义与 Faster R-CNN 保持一致，但是额外增加了一个 L_{mask}。假设需要预测的类别总数为 K，那么对于每一个 ROI，Mask 分支最终输出的 Mask 维度为 $K\times m\times m$，其中包含了 K 个大小为 $m\times m$ 的二值掩码，这样每一个输出的 Mask 都具有 K 个类别。在计算损失时，并非每一个类别的输出都要拿出来进行计算，而是根据该像素属于的具体类别，仅使用该类别的 Mask 输出来计算损失，这样就避免了类间的竞争，实际实验结果表明这种方式可以提高实例分割的效果。

1.2　基于 Mask 的边缘滤波优化

虽然使用 Mask R-CNN 网络已经可以得到比较好的分割结果，但是在实际应用时仍然会经常出现对目标物体分割不够精确的问题，一旦不能够正确地将目标对象从图像中分割出来，就会严重影响后续双目立体匹配的精度。本文在使用 Mask R-CNN 完成分割后，进一步基于得到的图像 Mask 引入基于 Mask 的边缘轮廓提取与滤波优化操作。

当图像某一区域出现了局部的不连续性时就代表到达了图像边缘，例如灰度、颜色、纹理等特征出现了突变。Canny 是目前效果最好的边缘提取算法，为了能够最大化边缘提取效果，提高 Mask 精度，本文引入 Canny 算子进行边缘提取。

Canny 算子求图像边缘点的具体计算过程如下：

首先将图像转为灰度图。Canny 算子仅能够对单通道灰度图进行处理，因此这一步是必要的预处理步骤。较常使用的灰度化公式为：

$$Gray=0.299R+0.587G+0.114B \tag{2}$$

然后经过高斯滤波平滑处理、梯度值和方向计算、非极大值抑制和滞后阈值处理后提取 Mask 边缘轮廓，具体流程如下：

(1)高斯滤波平滑处理

我们希望输入的图像是完美无噪声的，但是现实情况下由于所使用的拍摄工具、环境干扰等各种外部因素的影响，所采集的图像信息都带有大量噪声。通过使用高斯滤波，可以将图像中的噪声点去除。

使用高斯分布函数的二维形式，来生成一个大小为 $n\times n$ 的高斯滤波矩阵：

$$g(x,y)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-\frac{x^2+y^2}{2\times\sigma^2}} \tag{3}$$

以 3×3 的高斯核为例：

$$G_\sigma(x,y)=\begin{bmatrix} g(0,0) & g(0,1) & g(0,2)\\ g(1,0) & g(1,1) & g(1,2)\\ g(2,0) & g(2,1) & g(2,2)\end{bmatrix} \tag{4}$$

计算后就得到了一个高斯核，之后直接进行卷积操作即可。经过卷积计算后所得到的新图像 I_{out} 在点 $p(x,y)$ 处所对应的灰度值为：

$$p_{\text{new}}(x,y)=(G_\sigma * P_{x,y})(x,y) \tag{5}$$

(2)计算梯度值和方向

使用一阶函数计算图像在 x 和 y 轴一共 4 个方向上的梯度，通常使用 Sobel 算子作为梯度算子：

横向：

$$S_x=\begin{bmatrix} -1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \tag{6}$$

纵向：

$$S_y=\begin{bmatrix} 1 & 2 & 1\\ 0 & 0 & 0\\ -1 & -2 & -1\end{bmatrix} \tag{7}$$

对于输入的图像 $I(x,y)$：

$$I(x,y)=\begin{bmatrix} a_0 & a_1 & a_2\\ a_3 & c & a_4\\ a_5 & a_6 & a_7\end{bmatrix} \tag{8}$$

点 c 在横向和纵向上的梯度为：

$$\begin{cases} g_x=(a_2+2\times a_4+a_7)-(a_0+2\times a_3+a_5)\\ g_y=(a_2+2\times a_4+a_7)-(a_0+2\times a_3+a_5)\end{cases} \tag{9}$$

则在点 $c(x,y)$ 处的梯度幅度为：

$$G_{c(x,y)}=\sqrt{g_x^2+g_y^2} \tag{10}$$

点 $c(x,y)$ 处的梯度方向为：

$$\theta=\arctan\left(\frac{g_y}{g_x}\right) \tag{11}$$

(3)非极大值抑制

即使某一点处的梯度幅值很大，也并不能确定这一点就是边缘点，一个边缘点应该是局部的最大值，非极大值抑制就是去除那些非极大值数据作为边缘的可能性。将点 $c(x,y)$ 的梯度与其梯度方向上的点进行比较，如果不是局部最大值，那就说明这不是一个边缘点，则直接去除。

(4)滞后阈值

使用两个阈值来对图像进行筛选。具体来说，设定一个高阈值和一个低阈值，如果一个点的梯度幅值小于低阈值则说明不是边缘点；如果高于低阈值但小于高阈值，则认为是一个弱边缘点；如果该点的梯度幅值大于高阈值，则是一个强边缘点。首先在高阈值点组成的图像中尝试将边缘点连接形成轮廓，如果存在不闭合的边缘轮廓，则在断点附近加入合适的低阈值点，直到整个边缘闭合为止。

使用 Canny 算子处理流程提取出的 Mask 边缘轮廓作为初值，然后基于 RGB 图像对 Mask 轮廓进行优化修正，从而得到更加精细贴合目标的图像 Mask。实现方法如下：

遍历轮廓线上所有像素，以当前像素为中心开窗口计算窗口内像素 RGB 与当前像素 RGB 差值绝对值的和，判断差值和是否满足阈值，差值和大于阈值表示 Mask 边缘合理，差值和小于阈值表示 Mask 边缘不合理。对于不满足差值阈值的像素，遍历其窗口邻近像素。重复这个步骤直至边缘合理为止，再使用合理的像素替换 Mask 轮廓线像素。具体流程如算法 1 所示。

算法 1 Mask 轮廓边缘优化流程算法

1. Mask 获取：
   1.1. 基于 Mask R-CNN 网络提取 RGB 图像目标 Mask
   1.2. 基于 Canny 算子提取 Mask 轮廓线
2. 轮廓线插值：
   2.1. 根据 Canny 算子提取 Mask 轮廓线进行加权插值，使其更加平滑
3. 遍历轮廓线每个像素：
   3.1. 以当前像素为中心开窗口计算窗口内像素 RGB 与当前像素 RGB 差值绝对值的和
   3.2. 判断差值和是否满足阈值，差值和大于阈值表示 Mask 边缘合理，差值和小于阈值表示 Mask 边缘不合理
   3.3. 对不满足差值阈值的像素，遍历该像素窗口邻近像素：重复步骤 3.1—3.2 直至边缘判定合理为止，并使用最终合理的像素轮廓替换 Mask 轮廓线像素
4. 保存优化后的 Mask

图 2(a)为 Mask R-CNN 直接提取的车辆 Mask，图 2(b)为在 Mask R-CNN 提取车辆 Mask 经过 RGB 引导优化后的 Mask 结果。可以看到样例左边小车的边缘在 RGB 引导优化后更加平滑，样例右边小车的车头直接使用 Mask R-CNN 并没有提取出来，在经过 RGB 引导后就更加完整。

(a) (b)

图 2 基于 RGB 引导的 Mask 边缘优化

Fig. 2 Mask edge optimization based on RGB guidance

1.3 基于深度学习的双目点云重建

本文实现三维点云重建方法的整体框架是基于深度学习双目立体匹配网络 BGNet 构建，使用网络输出的视差图映射纹理点云，最后对得到的最初点云进行滤波与优化。

1.3.1 双边滤波加速的实时立体匹配网络(BGNet)

虽然近些年学者们对于双目立体匹配网络的研究已经有了非常显著的成果，但依然鲜有同时兼备实时性和高精度的网络出现。Xu 等基于可学习的双边网络[28]思想，提出了一种新的双边网络代价匹配体上采样模块(CUBG)，通过 CUBG 模块，可以高效地将输入的低分辨率代价空间转化为高分辨率下的代价空间。基于该模块，Xu 等设计了一个高精度的实时双目立体匹配网络 BGNet，网络的整体结构如图 3 所示。

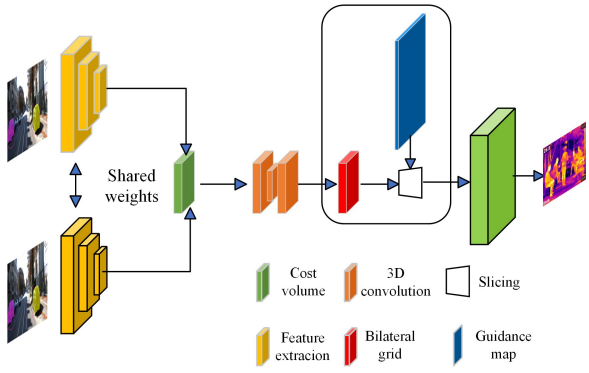

图 3 BGNet 网络结构

Fig. 3 BGNet network structure

图 3 中，网络包含 4 个主要模块：特征提取模块(使用一个类 ResNet[29]网络来提取图像特征)、代价空间聚合模块(通过 3D 卷积构建图像在低分辨率下的代价空间)、代价空间上采样(即 CUBG)和视差优化模块。

(1)特征提取

文中使用一个类 ResNet 结构来提取图像特征，在前三层使用 3 个 3×3 的卷积，用于对输入图像下采样；随后，通过 4 层残差层，生成低分辨率的一元特征；最后，连接这些输出的一元特征，形成一个具有 352 个通道的特征图。

(2)代价空间聚合

这里作者直接使用了 GwcNet[30]的分组相关代价空间结构，然后使用两个 3D 卷积进一步降低代价空间通道数，再通过一个类 U-Net[30]的网络结构进行代价聚合，最后输出一个 1/8 分辨率的代价空间。

(3)代价空间上采样(CUBG)

如图 4 所示，将一个低分辨率的代价空间 C_L 和高分辨率的图像特征作为 CUBG 模块的输入，最终得到一个高分辨率的代价空间。

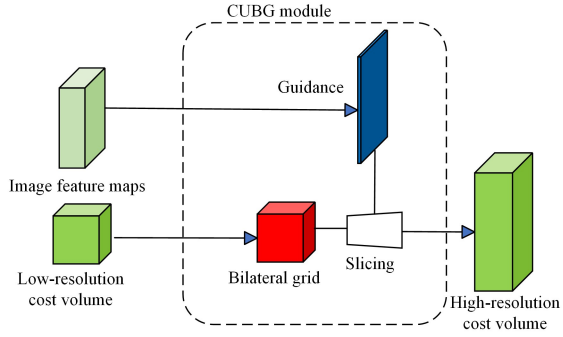

图 4 基于双边网格的上采样模块

Fig. 4 Upsampling module based on bilateral grid

设输入的是一个图像宽度 x、高度 y、视差范围为 d、通道数为 c 的代价空间 $C_L(x,y,d,c)$，使用一个大小为 3×3×3 的 3D 卷积将 C_L 转换为双边网络 $B[x,y,d,g]$，这里的 g 表示 B 的通道数。然后使用 slicing 操作对 B 进行上采样，得到维度为 $[W,H,D_{\max}]$ 的高分辨率代价空间。slicing 操作可以用式(12)来表示：

$$C_H(x,y,d)=B(sx,sy,sd,s_gG(x,y)) \tag{12}$$

其中，s 表示两个不同分辨率代价空间高度或宽度的比值，取值范围在(0,1)之间；$G(x,y)$为引导图特征。

(4)视差优化

使用 GCNet[30] 网络中提出的 softargmin 函数进行回归视差预测：

$$D_{\text{pred}}(x,y)=\sum_{d=0}^{D_{\max}} d\times\text{softmax}(C_H(x,y,d)) \tag{13}$$

(5)损失函数

如式(14)所示，BGNet 的损失函数可以表示为：

$$L=\sum_p \mathcal{L}(D_{\text{pred}}(p)-D_{gt}(p)) \tag{14}$$

$$\mathcal{L}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{otherwise}\end{cases} \tag{15}$$

其中，$D_{gt}(p)$为图像 p 的真实视差。

1.3.2 视差图映射纹理点云

在实时生成视差图后，需要基于双目相机模型参数解算得到深度图和三维点云。视差图转化为深度图可直接通过基线、焦距的视差公式转换，设点 $p(x,y)$深度为 D，视差为 d，基线长度为 B，相机的焦距为 f，则深度 D 的计算公式为：

$$D=\frac{Bf}{d+(x_{0r}-x_{0l})} \tag{16}$$

其中，x_{0l}和x_{0r}分别表示左右两张图像上主点的纵坐标，由于之前已经对两台相机做了极线矫正，故上式可以简化为：

$$D=\frac{Bf}{d} \tag{17}$$

得到深度图后，再将深度图转化为点云即可，对于点 $p(x,y)$在三维空间上的对应点 $p'(x,y,z)$，其坐标计算公式为：

$$\begin{cases}x=\dfrac{D(x-x_{0l})}{f}\\ y=\dfrac{D(y-y_{0l})}{f}\\ z=D\end{cases} \tag{18}$$

图 5 为由视差图生成的点云结果。

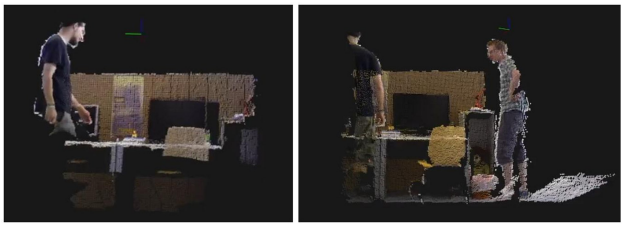

图 5 由视差图转换得到的三维点云

Fig. 5 3D point cloud obtained from disparity map conversion

1.3.3 点云滤波与优化

直接生成的点云中往往会因为不可避免的误差而存在大量散列点、孤立点。在对点云做进一步处理之前，只有将噪声点、离群点、孔洞、数据压缩等按照后续处理定制，才能够更好地进行配准、特征提取、曲面重建、可视化等后续应用处理。这一步需要根据实际任务的不同，灵活选择合适的滤波处理算法，如双边滤波、高斯滤波、条件滤波、直通滤波、基于随机采样一致性滤波等。

对于如何选择点云滤波的具体方案，PCL 库中总结了几种需要进行点云滤波处理情况，这几种情况分别如下：

(1)点云数据密度不规则需要平滑；

(2)因为遮挡等问题造成离群点需要去除；

(3)大量数据需要下采样；

(4)噪声数据需要去除。

对应的方案如下：

(1)按照给定的规则限制过滤去除点；

(2)通过常用滤波算法修改点的部分属性；

(3)对数据进行下采样。

(4)通过双边滤波算法进行噪声去除。

分别对 PCL 中几种常用的滤波算法原理作详细描述，通过大量实验探究每种算法在设置不同的阈值后的滤波效果及运行效率，得出如下结论：

首先，在对整体点云中的离群点滤波处理时，直通滤波与条件滤波效果不佳，但对筛选指定方向范围内的点云，其执行效率较高，适用于数据量大的点云数据集。

其次，统计滤波和半径滤波都能对离群点进行有效移除，二者的去噪效果较为接近，且设置阈值较为简单，但若点云数据量增大，其运行时间都将成倍增加，因此可根据实际情况，对数据进行降采样(如体素滤波)，来提升运行效率。

2 基于 GPU 加速的动态目标三维检测

为了对实时三维重建得到的点云实现动态目标逐帧监测和感知，本文提出了一种基于 GPU 加速的邻近帧点云的实时动态目标感知，针对三维点云数据量庞大、计算复杂的问题，引入 CUDA 技术将任务划分，交由 GPU 并行计算来优化运算性能，随后对所使用的动态目标感知算法进行实验优化，最后提出一种二三维一体化的动态目标实施感知技术。

2.1 GPU 软硬件架构

GPU(Graphics Processing Unit)通常作为计算机上的图形处理器，GPU 就是为了最大化并行计算能力而出现的，如图 6 所示，CPU 往往配备了大量的缓存结构和控制器单元来存储数据和控制指令，因此 CPU 在需要处理逻辑控制、快速实时相应时更具有优势，但是无法应对数据量较为庞大的任务；相反，GPU 本身的结构就较为简单，GPU 把大量的晶体管都用于算术逻辑单元，从而最大程度地提升计算速度和数据吞吐量，通过大量线程并行和提升线程的切换速度来掩盖存储器数据的访问时延，适用于处理计算密度高、数据类型统一、逻辑分支少的任务。

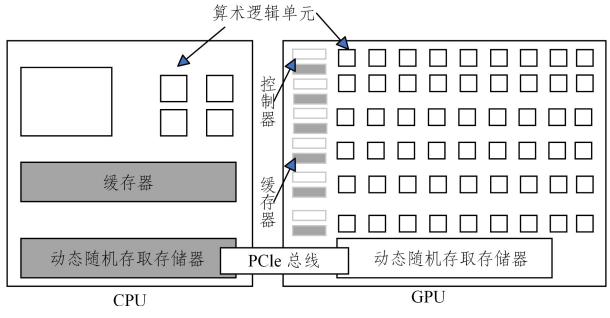

图 6 CPU 和 GPU 架构特点

Fig. 6 CPU and GPU architecture characteristics

通过使用 CUDA 将大型计算任务转交给 GPU 并行执行，可以加速应用的运行速度，并且能够执行因受 CPU 性能限制而无法执行的计算。CUDA 程序代码分为主机(host)端和设备(device)端两部分，主机端代码运行在 CPU 上，主要负责调控设备端代码的调度与运行；设备端代码运行在 GPU 上，主要负责执行设定好的计算任务。运行在 GPU 上的设备端代码一般称为核函数(kernel)。CUDA 程序在执行时，首先需要运行 CPU 上的 host 代码，之后所有 kernel 的启动、调用，以及分配的线程数都由 CPU 来控制。如图 7 所示，当一个核函数启动之后，会经历在 GPU 上生成网格(grid)，由网格划分为线程快(block)，最后划分为线程(thread)的整个过程。在 GPU 并行处理内部的数据时，同一个核函数内的不同 block 是异步执行的，不同 thread 之间是并行执行的，所有的运算过程之间不存在任何的依赖关系，它们相互独立，互不干扰，从而最大化 CUDA 程序的运算效率。

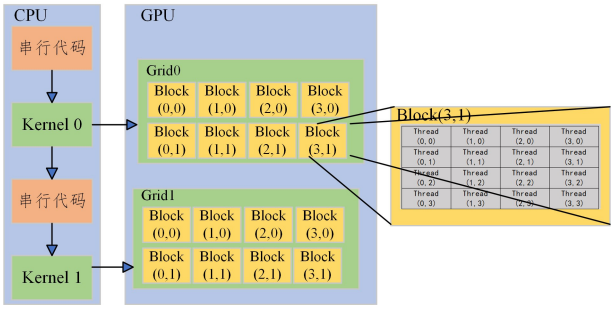

图 7 CUDA 线程层次结构

Fig. 7 CUDA thread hierarchy

2.2 基于邻近帧点云的动态目标检测

本文基于 GPU 强大的并行计算优势，提出一种基于邻近帧点云的动态目标检测方法，目的是通过输入相邻帧间的视差图，快速找到其中动态目标对应的三维点云，从而定位场景中的动态目标。

假设移动目标 T 在场景中上一帧的对应点云为 $P=\{p_1,p_2,\cdots,p_m\}$，当前帧中的对应点云为 $Q=\{q_1,q_2,\cdots,q_m\}$，那么如果目标 T 在两帧点云中存在动态的运动关系，则目标对应的点云 P 和 Q 在两帧之间一定存在一组矩阵刚性变化关系，R 和 t 分别是对应点云之间的旋转平移关系。那么对于 P 与 Q 上的每一个对应点 p_i 和 q_i，如果能够找到这样一组 R 和 t，使得 P 经过矩阵变换后得到的新点云与 Q 相减，这个残差越小，则证明得到的变化矩阵越精确。我们定义第 i 对点的误差项为：

$$e_i=q_i-(Rp_i+t) \tag{19}$$

则整体误差可以表示为：

$$J=\Sigma \| q_i-(Rp_i+T) \|^2 \tag{20}$$

显然当上式中的 J 取到了 0，或接近于 0 时，就可以认为找到了移动目标 T 在前后两帧中的对应点云以及它们之间的运动关系。那么要做的就是对 R 和 t 进行极值求解。我们构建最小二乘问题：

$$\min_{R,t} J=\frac{1}{2}\sum_{i=1}^{n} \| (q_i-(Rp_i+t)) \|_2^2 \tag{21}$$

然后尝试对式(21)求解，定义两组点云的质心：

$$q=\frac{1}{n}\sum_{i=1}^{n}(q_i),p=\frac{1}{n}\sum_{i=1}^{n}(p_i) \tag{22}$$

则误差函数可以处理为：

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n} \| q_i-(Rp_i+t) \|^2\\&=\frac{1}{2}\sum_{i=1}^{n} \| q_i-Rp_i-t-q+Rp+q-Rp \|^2\\&=\frac{1}{2}\sum_{i=1}^{n} \| (q_i-q-R(p_i-p))+(q-Rp-t) \|^2\\&=\frac{1}{2}\sum_{i=1}^{n}(\| q_i-q-R(p_i-p) \|^2+ \| q-Rp-t \|^2+\\&\quad 2(q_i-q-R(p_i-p))^{\mathrm{T}}(q-Rp-t))\end{aligned}$$

注意到上式中 $(q_i-q-R(p_i-p))^{\mathrm{T}}(q-Rp-t))$ 在求和后为 0，因此可以简化为：

$$\min_{R,t} J=\frac{1}{2}\sum_{i=1}^{n} \| q_i-q-R(p_i-p) \|^2+ \| q-Rp-t \|^2 \tag{23}$$

等式右边的两项中，可以发现左项只与旋转矩阵 R 有关，而右项虽然同时包含 R 和 t，但其中的 q 和 p 都为已知量，故只要得到了 R 也就能够计算出 t。于是整个问题变成了求解式(24)：

$$\begin{cases}q'=q_i-q,p'=p_i-p\\R^*=\arg\min_R \frac{1}{2}\sum_{i=1}^{n} \| q'-Rp' \|^2\\t^*=q-Rp\end{cases} \tag{24}$$

其中，q'，p' 表示 P，Q 上每个点的去质心坐标。

通过双目立体匹配得到的视差图与当前场景的深度图存在直接对应关系，而场景点云正是由深度图转换而来的，因此基于邻近帧点云的动态目标检测等价于基于邻近帧深度图像的动态目标检测。

而逐像素地进行邻近帧深度图的差异值检测效率低下且较为耗时，多数计算工作都只有少量的 CPU 与 GPU 时间消耗，最大的时间瓶颈就是基于大量深度图像素的逐帧比对。考虑到 GPU 本身具有极强的并行运算能力，本文设计了基于 GPU 框架并行的邻近帧点云目标检测算法，设计的 GPU 分块大小为 16×16，使用共享内存进行点云差异值并行计算，大大提升了动态目标检测效率。

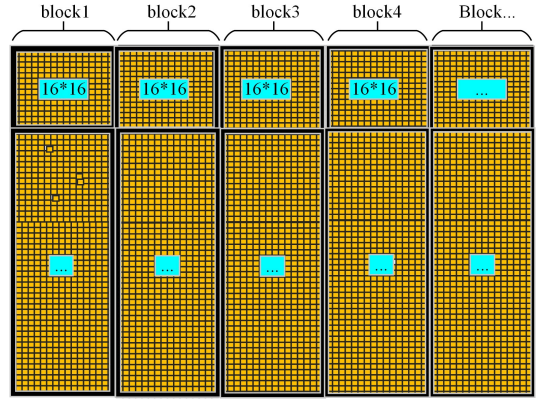

图 8 CUDA 加速视差图点云处理

Fig. 8 CUDA accelerated parallax map point cloud processing

3 实验与分析

3.1 实验数据与环境

实验软硬件环境如表 1 所列。

表 1　实验环境

Table 1　Experimental environment

软/硬件	型号/版本号
中央处理器	Intel(R)Core(TM) i9-10900KCPU @3.70GHz
图形处理器	NVIDIAGeForceRTX3090
显存	24GB
内存	32GB
操作系统	Windows11
Python 版本	3.8

本文实验部分使用慕尼黑工业大学的 TUM 数据集，数据集包含使用 Microsoft Kinect 传感器记录的多个不同室内场景的动态监测数据。数据集提供了 845 张分辨率为 640×1 480、帧率为 30 Hz 的高动态范围光度校准图像，以及相机运动的真实轨迹深度图像。如图 9 所示。

图 9　TUM 数据集

Fig. 9　TUM dataset

3.2　实验结果分析

本文使用动态数据集中的彩色图像进行两两邻近帧图像的立体匹配关系计算。针对每一对邻近帧，首先使用 Mask R-CNN 进行目标图像分割提取 Mask。

使用提取出的边缘轮廓对 Mask R-CNN 输出的图像 Mask 进行优化和再修正，从而得到更加精细贴合目标的图像 Mask。优化后的 Mask 结果如图 10 所示。

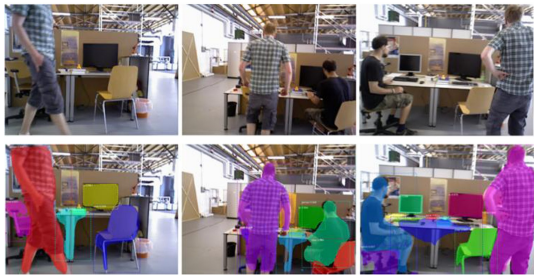

图 10　Mask R-CNN 分割结果

Fig. 10　Mask R-CNN segmentation results

从图 10 可以看到，使用 Mask R-CNN 获得的图像掩模初值虽然已有不错的质量，但对物体轮廓分割不够精细，通过 RGB 引导的 Mask 边缘滤波优化后得到的图像掩模能够更加贴合目标的边缘轮廓，如图 11 所示。

图 11　Mask 优化结果

Fig. 11　Mask optimization results

图 12 为双目立体匹配得到的视差图，映射到三维空间后生成点云，对于运动的人体能够实时有效地重建出目标的三维点云，并得到动态人体运动的轨迹数据，如图 13 所示。

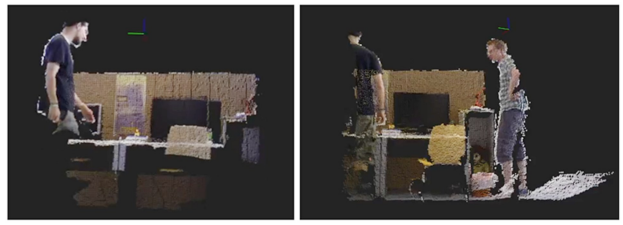

图 12　三维点云实时显示结果

Fig. 12　Real time display results of 3D point cloud

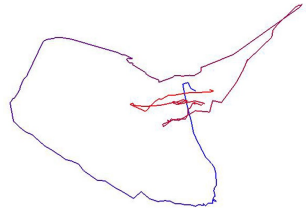

图 13　动态人体运动轨迹

Fig. 13　Dynamic human motion trajectory

结束语　本文结合各领域中的实际场景提出了一种基于双目估计的动态场景三维感知技术，在进行快速三维重建的同时对动态目标进行监测与感知。为了满足双目立体匹配算法对高精度的要求，本文提出了一种基于图像分割的 Mask 滤波与优化方案，用于提高双目立体匹配效率并提升视场点云重建精度。首先使用深度学习网络进行二维场景的实时分割与提取，得到剔除指定背景的目标 Mask，借助 Canny 算法提取精细的图像边缘轮廓，再对 Mask 轮廓边缘进行 RGB 颜色引导的边缘滤波优化，使图像 Mask 轮廓能够最大程度与目标对象拟合。基于双目立体几何深度学习网络进行实时目标点云生成，并提出基于 GPU 加速的邻近帧点云的实时动态目标感知(检测与识别)算法，针对三维点云数据量庞大、计算复杂的问题，通过引入 CUDA 技术，借助 GPU 并行计算从而大幅提高了邻近帧点云算法的计算速度。此外，本文提出的二三维一体化的动态目标实时感知技术可以与多种技术混合使用，从而获得更加全面、立体的环境空间信息，具有很好的应用前景。

参 考 文 献

[1] FANG L P，HE H J，ZHOU G M. A Review of Object Detection Algorithm Research [J]. Computer Engineering and Applications，2018，54(13)：11-18，33.

[2] WU Q，WANG T，WANG H W，et al. A Review of Modern Intelligent Video Surveillance Research [J]. Computer Application Research，2016，33(6)：1601-1606.

[3] ZHANG G Y，XIANG H，ZHAO Y. A review of research on computer vision based autonomous driving algorithms [J]. Journal of Guizhou Normal University，2016，32(6)：1674-7798.

[4] LI Q H，LONG X F，NONG Z L，et al. Clinical Application of Digital Medicine 3D Reconstruction Technology in Closed Abdominal Injury in Children [J]. Chinese and Foreign Medical Research，2021，19(3)：191-193.

[5] MARR D，POGGIO T. A computational theory of human stereo

vision[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1979, 204(1156): 301-328.
[6] BARNARD S T, FISCHLER M A. Computational stereo [J]. ACM Computing Surveys(CSUR), 1982, 14(4): 553-572.
[7] ZHANG Y W, HU K, WANG P S. A Review of 3D Reconstruction Algorithm Research [J]. Nanjing Information Technology Journal of Cheng University (Natural Science Edition), 2020, 12(5): 75-83.
[8] YOON K J, KWEON I S. Adaptive support-weight approach for correspondence search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 650-656.
[9] GERRITS M, BEKAERT P. Local stereo matching with segmentation-based outlier rejection[C]// The 3rd Canadian Conference on Computer and Robot Vision (CRV'06). IEEE, 2006: 66-66.
[10] LASKOWSKI Ł. A novel hybrid-maximum neural network in stereo-matching process [J]. Neural Computing and Applications, 2013, 23(7): 2435-2450.
[11] ZBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. J. Mach. Learn. Res., 2016, 17(1): 2287-2318.
[12] MAYER N, ILG E, HAUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4040-4048.
[13] PANG J, SUN W, REN J S J, et al. Cascade residual learning: A two-stage convolutional neural network for stereo matching [C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 887-895.
[14] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to-end learning of geometry and context for deep stereo regression[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 66-75.
[15] CHANG J R, CHEN Y S. Pyramid stereo matching network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5410-5418.
[16] TONIONI A, TOSI F, POGGI M, et al. Real-time self-adaptive deep stereo[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 195-204.
[17] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Image Net classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc., 2012: 1097-1105.
[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[19] HUANG G, LIU Z, WEINBERGER K Q, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[20] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]// Advances in Neural Information Processing Systems. 2014: 2366-2374.
[21] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2650-2658.
[22] SHELHAMER E, BARRON J T, DARRELL T. Scene intrinsics and depth from a single image[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 37-44.
[23] FU H, GONG M, WANG C, et al. Deep ordinal regression network for monocular depth estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.
[24] WOFK D, MA F, YANG T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems[C]// 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 6101-6108.
[25] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.
[26] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890.
[27] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv: 1412. 7062, 2014.
[28] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.
[29] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv: 1706. 05587, 2017.
[30] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 801-818.

HE Weilong, born in 1993, master, lecturer. His main research interests include photogrammetry, remote sensing technology, and computer vision.

SU Lingli, born in 1992, master, lecturer. Her main research interests include municipal engineering and computer vision.