

**硕 士 研 究 生 读 书 报 告**



题目 《即时三维摄影》读书报告

作者姓名 王业

作者学号 21951025

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2019 年 12 月

Instant 3D Photography

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Ye Wang

Zhejiang University, P.R. China

2019

摘要

虚拟现实（VR）是一种近年来发展迅速的新兴技术。通过计算机模拟，VR可以产生一个三维的虚拟世界，让用户感觉身临其境并能即时、无限制地观察三维空间内的事物。但目前VR所生成的虚拟内容大部分都是经过合成的，需要由专业人士来创作。

为了让普通用户也能利用VR技术来捕捉和分享自己周围的真实世界环境，本文提出了一种只需要一台双摄手机，就可以将所拍摄的照片即时转换成3D全景图的算法。该算法使用通过双摄手机捕获的一系列颜色和深度对齐的图像序列作为输入，经过可变形深度对齐、无缝拼接、多层网格生成等步骤，实现了可以在任何图形引擎中进行渲染、并具有运动视察的3D全景图的生成。本文提出了深度对齐的优化算法及加速图像拼接的算法，在几十秒内就可以完成一张十分精细的3D全景图的生成。

**关键词**：3D重建、图像拼接、实时渲染、虚拟现实

Abstract

Virtual Reality (VR) is an emerging technology which has developed rapidly recent years. Through computer simulation, VR can generate a three-dimensional virtual world, allowing users to feel immersive, and can instantly and unrestrictedly observe things in three-dimensional space. However, most content that is consumed in VR today is synthetic and needs to be created by professionals.

To allow ordinary users to use VR technology to capture and share the real-life environment around them, this paper presents an algorithm that can convert photos directly into 3D panoramas with dual camera cell phones. The algorithm uses a series of aligned color-and-depth image pairs captured by a dual-camera phone as input, generating 3D panoramas with motion inspections which can be rendered with any standard graphics engines through deformable depth alignment, seamless stitching and multi-layer mesh generation. This paper presents an optimization algorithm for depth map alignment along with an algorithm for accelerating image stitching, which can complete a detailed 3D panorama generation in a few tens of seconds.

**Keywords：**3D Reconstruction, Image-stitching, Real-time rendering, Virtual Reality

1介绍

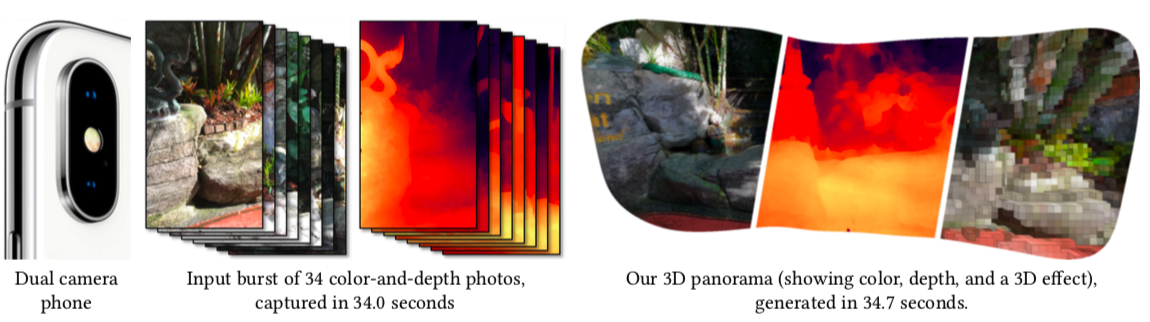
随着虚拟现实技术的迅速发展，越来越多的人可以在计算机模拟的虚拟场景中获得身临其境的体验——只需要一台头戴式VR设备，就可以畅游在具有真实感的的虚拟3D环境中。然而普通用户所能体验到的场景只能由专业人士创作得到，自己是无法将身边的景物转化到VR场景中的。为了让所有人都能充分利用VR技术来进行创作，伦敦大学学院的Peter Hedman和Facebook计算摄影部门的研究科学家Johannes Kopf合作完成了这篇论文，在论文中实现了即时3D摄影技术：该算法可以从一组照片中快速重建出具有景深和视差的三维全景图，并且这一系列过程不需要复杂的硬件，在普通的双摄手机上就可以完成。

文章中提出的算法处理速度很快，大约每秒钟就可以处理一个输入图像。该算法的输出是一张3D全景图，或者说是一张可以用标准图形引擎渲染的具有纹理的多层3D网格。这样一张3D全景图可以在具有双眼和头部运动视差的VR中进行观察，也可以在具有视差效果的手机和显示器中查看。

该算法所遇到的主要难点和针对难点所作出的主要技术改进有以下两点：

1. 由于双摄手机的两个摄像头基线非常窄，导致深度估计具有很高的不确定性，并且需要较强的边缘感知滤波来平滑结果图中的噪声。这种深度的不确定性会导致深度图中的低频误差，从而使算法无法使用简单的全局变换来进行深度对齐。该算法提出了一种新颖的优化方法：通过恢复相机的朝向和方位及求解深度值的空间变化调整场联合来对齐深度图。该方法能够很好地对齐严重降质的深度图。
2. 现有的使用离散优化拼接图像的方法很慢。为了加快图像拼接的速度，该算法使用经过仔细设计的数据项和高质量的深度对齐，从而避免了对标签平滑进行优化的需要。同时，在以深度引导的边缘感知方式过滤数据项后，该算法以独立优化每个像素标签来代替对每个图像标签的优化。这种优化方法实现了超过两个数量级的加速。

文章作者在各种场景中应用了该算法，包括室内外、城市景色和自然环境。实验结果表明该算法具有很高的鲁棒性，从各种照片中都能生成良好的对齐深度图和逼真的拼接全景图，例如图1-1所示。



**图1-1 左图：使用双摄手机以每秒一幅图像的速率拍摄的一组颜色和深度图像序列。右图：使用文章中算法生成的3D全景图，生成速度与拍摄速度大致相同。全景图非常精细，可以在VR中使用双眼和运动视差进行查看。**

**2 算法**

**2.1 现有算法的缺陷**

现在市场上已经有专用的硬件设备可以轻松捕捉360°× 180°全景图了。但这种全景图如果应用到VR场景中，是无法产生双眼和运动视差的感受的，也就在很大程度上限制了图像的真实感。

为了弥补这个缺陷，有一些算法通过拼接左眼和右眼的全景图像对来实现双眼深度感知，这种方法被称为全向立体成像技术。但这种方法也有很多缺点，如渲染得到的场景因不是在线性透视投影下而会表现出扭曲和错误的立体视差。更重要的是，用这种方法生成的全景图不支持运动视差，也就是说当用户穿戴VR设备时，渲染的场景不会因用户头部的移动而发生改变。这极大地限制了深度感知，使用户不能在VR场景中产生沉浸感。

另一种代替从左右眼图像生成全景图的方法是在传统的全景图中增加深度信息。目前已经有算法能够从窄基线360°视频中构建具有深度信息的全景图，但重建得到的深度信息仍然不足以实现运动视差的效果。Lee等人使用深度信息来计算空间变化的3D投影表面，以此来补偿使用360°相机拼接图像时的视差。然而这种方法只能生成低分辨率的表面网格。

要改善由3D重建所得到的全景图的质量，可以使用一项被称为多视图立体重建（MVS）的技术。MVS算法被广泛应用于VR场景的3D重建中。使用这种算法，可以从360°相机拍摄的视频序列中获取到密集的物体表面点云。然而MVS方法仅适用于基线宽的摄像机，在本文所讨论的全景捕获场景中并不能产生令人满意的效果。在这种情况下，很难处理三角剖分的不确定性所导致的噪声重建和区域缺失。MVS方法通常也很慢，运行时间从几分钟到几个小时不等。

总而言之，以上这些为深度传感相机所设计的方法不都不适用于本文所讨论的窄基线重建数据。深度传感相机所拍摄的宽基线图像可以提供绝对比例并保证帧到帧的一致性，通常可以高精度地对齐其深度图。然而本文算法的输入是双摄手机所拍摄的窄基线图像，其深度图的质量很低，需要进行额外的对齐工作，这也是这篇文章中提出的算法所重点讨论的内容。

**2.2 算法概述**

文章中提出的3D全景图重建算法，总结起来可以分为以下四个步骤：

1. 捕获图像。算法的输入是一系列深度和颜色对齐的图像对，这些图像对通过普通的双摄像头手机内置的相机就可以拍摄得到。
2. 可变形深度对齐。由于相机的窄基线会导致三角剖分的不确定性，所得到的深度图是非常不精确的，并且不能使用全局变换来实现高质量的对齐。文中算法使用了一种新颖的优化方法：通过恢复相机的朝向和方位及求解深度值的空间变化调整场联合来对齐深度图。该方法能够对深度图进行形变，实现高质量的对齐效果。
3. 拼接。接下来，要将对齐的深度和颜色图像拼接成全景图像。通常来说，这是一个标签对齐问题，一般使用离散优化方法来完成。然而，使用如MRF求解器来优化标签平滑度是非常慢的。文中算法使用经过仔细设计的数据项和高质量的深度对齐（在步骤2中得到），从而避免了对标签平滑进行优化的需要。同时，在以深度引导的边缘感知方式过滤数据项后，该算法以独立优化每个像素标签来代替对每个图像标签的优化。这种优化方法实现的效果在视觉上和传统方法是相近的，并实现了超过两个数量级的加速。
4. 生成多层网格。最后一步，将在第三步中得到的全景图转化为一张可以在任何标准图形引擎中进行渲染的多层网格。通过在强深度边缘撕裂网格并将背面像素延伸到被遮挡的区域，就可以在被遮挡的区域产生新的颜色和深度值。最后，简化网格并计算纹理贴图。

**2.3 算法难点**

获取深度图并不是一件难事，因为深度图拍摄模式已经被集成到iPhone7中，其使用的是原生的iOS立体重建算法。只需要手持手机，将摄像头指向外方，同时任意向上、向下或侧身左右移动手机拍摄一系列照片即可。这里有一个细节在于，iPhone7在拍摄深度图时相机视野比较窄，需要捕获比其他相机更多的图像，大概包括20到200个图像。在拍摄的同时还需要使用IMU记录相机朝向的方位以供之后的对齐使用。

算法的第一个也是最重要的难点在于：对齐深度图。问题在于，图像是从不同的视角拍摄的，而深度图也是没有绝对比例的，例如，当深红色表示100英尺时，浅黄色并不表示10英尺，而且每张照片的比例都不相同。这意味着需要恢复相机的方向和位置，并以某种方式使这些图像的深度图在3D投影下对齐。作者尝试了传统的对齐方法，即对深度图使用全局变换：估算出每张深度图的变换系数，从而将归一化后的深度系数（0到1之间，如0.5）转化回绝对深度（如100英尺）。但经变换后得到的深度图效果却不是很好，如图2-1(a)所示，因为每张深度图都是由未知参数的曲线方程归一化得到的，优化器计算不出正确的变换系数，也就无法对齐它们。另一方面，由于基线过窄，三角剖分的不确定性导致立体深度图中出现低频误差，这也表示使用简单的全局变换是不可能正确地对齐这一系列深度图的。

针对这个问题，作者提出了通过估计空间变化的调整域来对深度图进行变形，从而使它们可以更好地对齐。算法使用了一些经过插值的网格来代替全局变换的仿射模型中的放缩和位移系数，并且通过减少相邻网格之间的差异值来增强结果的光滑度。实验结果如图2-1(c)所示，可以看到对齐后的深度图非常光滑并且没有浮子，表明这种可变形对齐算法可以实现高质量的对齐效果。



**图2-1 (a)使用全局变换仿射模型对齐后的深度图。(b)使用SFM点云重建后的深度图。(c)使用可变形仿射模型对齐后的深度图。可以看到可变形对齐算法生成的深度图效果最好。**

获得对齐的3D深度图后，下一步需要将这些深度图拼接为没有缝隙的全景图。这一步是算法的第二个难点：传统的图像拼接技术是通过标记图像中的标签来完成各图像的标签匹配，从而将一系列不完整的区域图像拼接成一张完成的全景图像。然而正如之前所说，对标签进行平滑处理需要大量的计算时间。为了加快这一步骤的处理速度，作者将这个问题转化成离散的标记问题：单独处理每个像素，而不是每个标签，也就是需要为每个像素从全景图中寻找对应的源图像，从而确定它的深度和颜色。算法根据一定的“惩罚项”计算出每个像素的数据项，并对得到的数据项进行优化。独立优化每个像素的速度虽然很快，但由于标签变化得非常频繁，会产生噪声。为了过滤所得到的数据项中的噪声使结果更平滑，算法提出了基于成本-体积的过滤器，它使用深度引导边缘感知过滤的方式来过滤数据项。结果如图2-2所示，可以看到虽然标签变化得很频繁，但拼接后得到的色彩图和深度图效果都非常好。



**图2-2 改进后的拼接算法得到的标签标记结果、色彩图和深度图**

算法的最后一步是将所得到的全景图转化为一张可被标准图形引擎渲染的三角网格。一种最简单的创建三角网格的办法是将所有像素点连接到与它们相邻的四个像素点。但是使用通过这种方法得到的网格来渲染全景图时，所得到的结果在观察的时候会出现如图2-3(a)这样的问题：当观察的视角发生变化时，在物体的强深度边缘处的网格会被拉伸。

为了解决这个问题，作者先检测出网格中的主要深度边缘并使用中值滤波器将柔和的边缘转化为阶梯边缘，如果相邻顶点之间的深度差距较大，就撕裂它们之间的连接。图2-3(b)是在深度边缘处撕裂顶点后得到的网格。这一步就可以将不同深度的物体隔离开来，使它们看起来处于不同的深度，并可以随着视角的变化产生透视的移动效果。但由于在视角发生变换时会观察到物体背面被遮挡的部分，还需要在深度边缘处生长网格以填补物体的背面网格。在生长网格的每次迭代中，对四个基本方向之一中缺少连接的每个顶点在此方向上生长并生成与其深度相同的新顶点，并将这个新顶点与其相邻的边缘顶点连接起来。经过多次迭代即可生成完整的网格。通过这个方法生成的网格，在深度边缘处平滑地延伸到被遮挡区域，如图2-3(c)所示。可以看到，现在当观察视角发生变化时，被遮挡区域不再是被拉伸的三角形，而是平滑的色彩和深度。



**图2-3 (a)原始方法生成的网格 (b)在深度边缘处撕裂顶点所得到的网格 (c)使用多层网格算法得到的网格。可以看到被遮挡区域被填充了平滑的色彩和深度内容。**

**3 分析与讨论**

作者将算法应用于大量的摄影图像中，包括了如白天和黑夜、室内和室外、城市和乡村等各种场合的景色。结果表明算法非常成功，在几乎所有图像中都有很好的表现。算法可以实现对深度图良好的对齐，并且在VR场景中可以产生逼真的双目视差和运动视差的观察效果。更重要的一点是，有些已有的3D重建系统（如微软研发的Casual 3D系统）虽然也可以产生这么逼真的结果，但其花费的时间要比本文所提出的算法慢上几个数量级：本文算法在几十秒内就可以完成一张全景图的生成，而Casual 3D处理同样一组输入则需要1-2个小时。

但在我看来，这个算法有一项不足之处：对于物体背面被遮挡的部分，本文的算法是使用漫反射修复的方法来进行填充，也就是合成与其相邻的像素点的颜色来作为背景色填充到被遮挡区域。这样，在视角变换进行观察时，所看到的物体背面的区域就会给人一种不真实感。解决这个问题的办法可能如下：通过卷积神经网络去猜测和填充图像的其余部分。例如，如果图像区域有头发，那么头发可能会继续存在。因此，它可以令人信服地重新构建这些纹理，并估计对象的形状、缩小间隙等。这样当改变视角时，就会观察到似乎是真实存在的画面。但由于训练卷积神经网络所花费的时间较长，并且GPU在这个过程中要扮演重要的角色，这和本文的目的和思想就不符了：本文的主要目的就是减短重建所需时间，并且在普通的CPU中就可以完成算法的大部分步骤，这也许就是为什么文章没有使用我所提到的这种方法的原因。

**4 小结**

通过阅读这篇文献，我深入了解到了一些关于VR技术和3D重建的知识，比如3D全景图的构建可以通过双眼观察到的两幅图像进行拼接得到，从而可以实现双眼视差，增强3D效果的真实感；但VR场景的构建不仅需要双眼视差，还需要运动视差，即当观察者的头部转动时，可以观察到360度各个视角下的全景图。这篇文章突破了原有3D重建算法的一些瓶颈，如运行时间过长、输入图像要求过高等，实现了“亲民”的算法，使人们仅仅使用手机就可以实时地将拍摄到的照片转化为VR虚拟场景。这种算法如果可以应用到各种社交网站和app如微博、朋友圈、FaceBook中，那么人们就可以实时拍摄和浏览具有VR感的3D图片，不再仅仅被约束在2D平面空间中了。期待这项技术能尽快被普及到我们的生活中。

参考文献

[1] Peter Hedman, Johannes Kopf. ACM Transactions on Graphics, Vol. 37, No. 4, Article 101. August 2018.

[2] Robert Anderson, David Gallup, Jonathan T. Barron, Janne Kontkanen, Noah Snavely, Carlos Hernandez Esteban, Sameer Agarwal, and Steven M. Seitz. 2016. Jump: Virtual Reality Video. ACM Transactions on Graphics 35, 6 (2016).

[3] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. 2017b. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. ACM Transactions on Graphics 2017 (TOG) 36, 3 (2017), Article no. 24.