基于语义交互的三维重建

史利民，郭复胜，高 伟，胡占义
（中国科学院自动化研究所模式识别国家重点实验室 北京 100190）
（slm_sx@126.com）

摘 要：三维重建是计算机视觉的核心问题.随着三维重建技术的发展，人们越来越认识到从图像到空间结构这种自底向上的重建方法，不管如何刻意设计和优化都很难达到对场景具有高层语义意义下的重建.基于此，提出一种基于高层语义交互的重建方法，在底层重建结果的基础上，利用自然语言交互进一步完善重建结果.最后通过 2 组简单实验验证了文中算法的可行性.

关键词：三维重建；语义交互；知识库
中图法分类号：TP391

3D Reconstruction via Semantic Interaction

Shi Limin, Guo Fusheng, Gao Wei, and Hu Zhanyi
(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract: 3D reconstruction is a key problem in computer vision. With the advances of 3D reconstruction techniques, people more and more realized that under the tradition paradigm purely from images to 3D structure, it is nearly impossible to recover the scene structure in a semantic sense no matter how deliberate design and optimization are employed. To this end, a semantic interaction based scene reconstruction approach is explored where automatically reconstructed noisy points or lines are upgraded through natural language understanding. The feasibility of our approach is illustrated by two simple experiments.

Key words: 3D reconstruction; semantic interaction; knowledge base

1 相关工作

基于图像的三维重建尽管有各种各样的方法，但基本过程不外乎是图像预处理、特征提取和描述、特征匹配、摄像机标定、空间形状重建（主要是点线重建）、整体优化等步骤，这种重建方法是一种自底向上的信息提取过程.经过 30 多年的研究，已有一系列关于三维重建理论、算法和系统的文献报道，特别是近期 Furukawa 等的关于多视立体重建的系列性工作[1-4]较以往工作有了很大的进步.但总体来说，很多报道中所谓的优良系统和算法的重建结果很难达到人类视觉感官的效果.而对室内场景重建这种现象表现得尤为明显.人们对日常生活中常见的物体，如桌子、椅子，重建的结果从量测的观点看误差也许并不大，但从感官会产生明显的差异，如桌面不平，

收稿日期：2010-09-08；修回日期：2010-11-19.基金项目：国家自然科学基金（60835003）；国家“八六三”高技术研究发展计划（2009AA012103）.史利民（1977—），男，博士研究生，主要研究方向为计算机视觉；郭复胜（1981—），男，博士研究生，主要研究方向为计算机视觉；高 伟（1980—），男，博士，助理研究员，主要研究方向为计算机视觉；胡占义（1961—），男，博士，研究员，博士生导师，主要研究方向为计算机视觉.

垂线不垂直等.尽管有些重建系统使用了共面性、垂直性和平行性等约束,但由于室内场景多种多样,将固定的一些规则引入到重建系统中很难得到好的效果.当这种自底向上的重建方法达到某种程度后,如果对其继续刻意地优化和完善已不可能再取得实质性进展,显得意义不大.这种自底向上的重建完成后,引入用户对场景的一些定性和定量的高层描述信息,可望有效地改善最终的场景重建结果.

在重建过程中,引入用户对场景描述的一些高层知识的困难在于如何方便地引入这些知识,以及如何利用这些知识来改善重建结果.因为没有任何人比用户对待重建的场景更了解,也没有任何人机交互方式比自然语言的交互形式更自然;所以本文研究基于语义交互的三维重建,它可以概括为当通过传统方法得到场景重建的(不完整且有噪声)点云和线段后,探索如何通过自然语言交互方式,并利用用户所提供的一些关于场景的定性或定量的知识来改善重建效果的途径和方法.

就我们所知,这种基于语义交互的重建方法在计算机视觉领域还没有相关报道.但下面的工作与本文研究内容有一些相关性:

1) 利用鼠标进行交互重建是最直接的交互方式,其交互的可靠性和准确性很高,是一种劳动密集型方法.该方法与其说是机器在进行三维重建,不如说是人在进行三维重建.事实上,长期以来,人们对三维重建研究的不懈努力就是希望探索自动重建的方法来改善这种劳动密集型方法的过程.

2) 基于领域知识的重建.这类方法一般针对一些特定的用途(如汽车零件)建立各种子模型,并在模型的指导下进行重建.从某种程度上说,本文方法与这类方法有一定的相似性.但本文方法与这类方法也有本质的不同:基于领域知识的重建方法将领域知识定义为一些通用模型,重建时从模型库中逐渐寻找合适的模型;而本文方法根据具体场景由用户直接提供自己对场景的一些理解信息,是在用户对特定场景提供高层语义指导下的重建.由于没有任何数据库中的知识能比用户对当前场景提供的知识更准确,所以至少从原理上来说,本文方法更具有提高三维重建质量的潜力和优势.

3) 基于上下文学习的方法.该方法目前是计算机视觉的一个热门话题,如ECCV2008最佳论文[5]、ICCV2009马尔奖[6]以及其他一些工作[7-10].基于上下文学习的方法通过学习一些物体与物体(或物体与环境)之间的关系,并利用这些学习到的关系来提高物体的识别率.这种在物体识别中学习上下文的思路同样可以用到三维重建,一个典型的例子是Saxena等从单幅图像学习三维结构的例子[11].但对不同场景均成立的上下文知识不可能很多,所以将上下文知识直接融入到自底向上重建过程的方法也不可能对重建效果有大的改善,特别是其普适性不可能高.

4) 基于邻域知识的方法.最典型的方法是带有平滑约束的能量优化方法,如图割法[12-13].这类方法可以说是基于底层普适知识的重建,不可能提供高层知识来对初始的三维重建结果提供有效的约束.

总之,目前文献中还没有类似本文拟探讨的基于语义交互的三维重建的报道;上面几方面的内容仅仅在原理上与本文工作有一些关联.需要指出的是,本文的目标仅仅是探索基于语义交互进行三维重建的可行性,并不意味着本文工作比其他工作好,也不意味着本文工作目前的状态已经具有实用性.从后面的实验中可以看出,本文工作仅仅是一种尝试性探索、是一项起步性工作.

另外需要指出的是,本文虽然研究用户语义交互下的三维重建,但并没有介绍计算机是如何理解用户"通过自然语言提供的交互信息"这个过程的.这主要是因为从当前语音认别的研究进展看,计算机理解用户的一些简单交互语言已不是一个困难的问题,并且这方面的内容也不是本文关心的主体.

2 基本思路

室内场景缺乏纹理的特性决定了单纯依靠特征匹配来重建必然只能得到非常稀疏的点云,不足以用来描述场景.另外,由于标定误差、匹配误差、重建误差等,使得有些重建出的点和线可能偏离真实的位置,在结构化的场景中这种偏差尤为明显,让人感觉"参差不齐".在这种情况下,要想得到比较令人满意的重建效果,在当前的重建算法基础上仅仅依靠底层的特征信息是很困难.于是,借助高层信息来辅助重建成为一个自然的选择,即结合底层特征信息和上层的感知信息来完成场景重建.这种利用高层信息的思路虽然曾被应用在基于激光扫描和立体图像对的稠密重建中[14-16],但相对这些方法得到的稠密数据而言,一般的视觉重建得到的点云要稀疏得多,所以难度更大,与本文方法有本质的不同.

由于用户通过自然语言交互是一种最直接和自然的交互方式,同时用户对待重建的特定场景能够

提供更准确和相关的场景知识，所以本文探索用户在自然语言高层语义交互下的三维重建方法.

虽然室内场景的单调纹理给传统的重建带来了很大困难，但从另一方面来讲，室内场景陈设的表面结构比较简单，即使较为复杂的物体也大多只包含相互垂直的 3 个方向的平面（如桌子、柜子等）. 而且，常见的室内物体结构基本固定，也就是说同一物体上不同位置的表面之间具有相对固定的关系（如平行、垂直、邻接等）. 这些特性决定了与室内场景相关的高层知识不会太复杂.

运用常用的点线匹配重建得到的稀疏点云和直线段都分布在物体的表面，特别是其中大部分都分布在物体的边缘部分，即在一定程度上能描述物体的基本框架，这些都为物体表面的定位提供了重要信息. 本文试图通过将底层得到的稀疏信息和用户提供的高层知识结合起来，共同完成室内场景中物体的表面重建.

3　本文算法流程和实现

本文算法的实验平台如图 1 所示，该平台中相机可以绕轴水平旋转，且旋转角度可以由控制器读取，相机的倾角以及离开旋转轴的距离也可以自由调节. 该平台的标定及点线重建过程见文献[17-18].

图 1　数据采集平台

本文算法主要包含底层重建和高层重建两大部分，如图 2 所示. 底层重建通过特征点、直线的匹配重建出空间点云和直线段，具体过程详见文献[17-18]. 需要指出的是，文献[17]中将重建点云的世界坐标系 z 轴与平台的旋转轴平行，故点云中地面上的点平行于 $x-y$ 平面，这一特点十分有利于地面的确定. 确定地面大致位置的直方图算法如下：

设 $P_{z\min}$ 和 $P_{z\max}$ 分别是点云（包含直线段）在 z 向上的最小值和最大值. 首先沿 z 轴在 $[P_{z\min}, P_{z\max}]$ 内以 $h=2$ cm 为步长将点云分割在 $\lceil (P_{z\max}-P_{z\min})/h \rceil$ 个 bins 内. 取含有点云个数大于阈值 S（本文实验取 $S=200$）的索引值最小的 bin 内的点，其 z 坐标平均值 z_0 确定了地面所在平面 $z=z_0$.

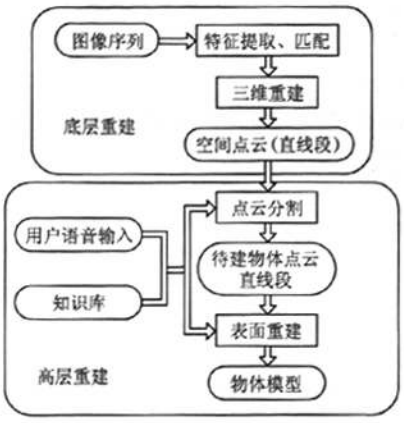


图 2　本文算法框架

为了高层处理处理方便，本文在底层重建结束后，通过直线图算法确定地面所在平面 $z=z_0$；然后删除 z 坐标小于 z_0+deta（本文中 $deta=5$ cm）的所有点和直线段，就得到了仅仅属于场景中各物体的点云和直线段. 将这些物体点云和直线段平移到以地面为 $x-y$ 平面的坐标系下，即作 z 向平移 $z=z-z_0$.

高层重建是在底层重建的基础上，利用用户所提供的场景信息结合知识库来进一步改善重建结果，以尽可能完整、正确地重建出场景中的物体. 假定用户对场景熟悉或者就在场景之中，从而能对要重建场景的基本信息有一些比较准确的估计或掌握. 理论上，只要重建出一个物体的所有表面，则该物体就被完整地重建出来了. 在纹理缺乏的室内场景中，通过底层重建只能得到稀疏的“点和线”，高层重建的目的是利用底层重建的点线信息结合人的高层感知信息共同完成物体的“面重建”，当然“面重建”的结果依赖于“点线重建”. 对于一个较为复杂的物体，要重建出它所有的面并非易事，特别是在存在遮挡的情况下. 本文方法充分利用已知信息，尽量多地重建出物体的主要面结构.

本文中关于高层重建算法的设计思路类似于人的识别过程. 比如人在识别一个桌子时，遵循的规则是逐个判断桌子的各个组成部分是否满足头脑中桌子的各种属性、约束，最后来决定其是否是桌子. 而本文算法是将人类关于桌子表面的知识放入到知识库，根据知识库约束来搜索对应于桌子表面各个部位的点云，并根据点云信息逐个重建出这些部位的表面，最终完成整个表面的重建. 可以看出，这是

具有一定智能的重建，故称之为高层重建，这部分是本文算法的核心，下面对此做具体的描述.

3.1 知识库

既然要依靠语义来进行面的重建，那么描述物体的知识库是必不可少的. 由于最终要重建的是物体的表面，故本文所指的一个物体的知识实际上指的是其表面知识，主要包含结构和属性 2 部分内容. 结构反映物体表面的组成，组成物体表面的基本元素称为基元. 本文设定所有组成物体表面结构的基元是平面或二次曲面. 图 3 所示为几个常见物体的表面结构(基元组成)图(本文假定圆形茶叶筒是圆柱体，柜子形状是长方体)，基元名称中的数字表示包含该基元的个数. 对于一个较为复杂的物体(如 L 形电脑桌)，其包含必要和可选 2 类基元. 必要基元是该物体必须含有的基元，可选基元为物体可能含有也可能不含有的基元. 属性反映物体及其组成基元的外形特征(如长、宽、高、位置、半径等参数)，以及物体组成元素之间的关系如邻接、平行、垂直等，其中只有所有数值属性是可以修改的；其他属性是物体的固有属性，反映了物体及各基元的固有特征和关系. 只要确定出每个基元的所有数值属性值，物体表面就可以完整重建出来.

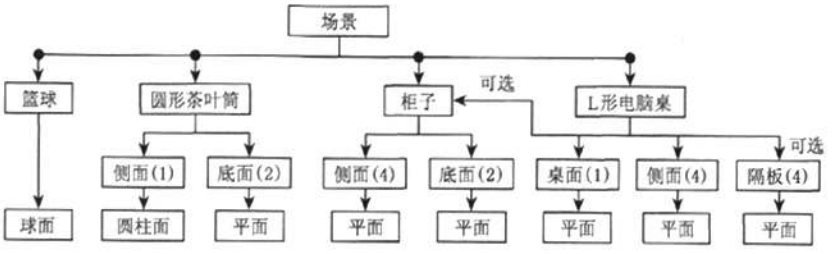
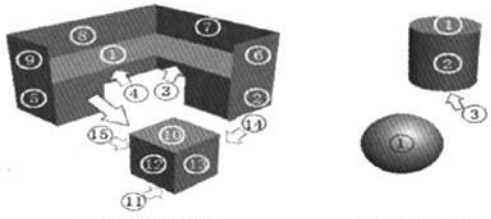


图 3 室内场景中一些物体的表面结构即组成基元

图 4，表 1，2 所示为一个简化的电脑桌和几个常见的基本几何体的表面结构图和属性表，属性表中数值属性的默认值是该物体通常情况下的取值或取值范围，用户可以通过输入具体的参数修改这些值，也可以使用这些默认值. 同一物体的不同基元可能具有相同的属性值，由于用户输入的参数只是对当前物体的相应属性提供了参考数据，并不一定精确符合物体，所以在高层重建过程中允许在一定范围内对其修正，本文实验中修正范围是参考值的 ±20%. 如一个面的输入参考宽度为100 cm，则重建该面的宽度允许在 80～120 cm 之间. 知识库的作用就是引导算法有目标地搜索各个组成基元，并确定其属性参数，逐个重建出基元，最终完成物体表面的重建.



a L形电脑桌模型 b a的基本几何体

图 4 一个简化的 L 形电脑桌模型和几个基本几何体模型

表 1 几种常见基本几何体属性表

名称	基元		属性	值
球体表面	必要	球面①	半径	12.5 cm
			球心	待定
圆柱体表面	必要	底面①②	形状	圆形
			半径	待定
			圆心	待定/待定
			位置及关系	①②平行，且距离为侧面高度
	必要	侧面③	形状	圆柱面
			高度	待定
			截面圆心及半径	同底面
			关系	垂直相交底面

表 2　L 形电脑桌属性表

名称		基元		属性	值
L 形电脑桌	必要	桌面①		形状	L 形状
				高度	80±20 cm
				长度	150±50 cm
				宽度	150±50 cm
				方向	平行地面
		侧面②③④⑤		形状	矩形
				高度	同桌高
				宽度	②⑤同桌面宽，③④同桌面长
				方向	垂直地面（桌面） ②④平行，③⑤平行 ②③互相垂直
				位置	②⑤在桌面两端 ③④在桌面外侧边缘，且交于 L 拐角处
	可选	隔板⑥⑦⑧⑨		高度	120±20 cm
				其他属性	同侧面
		柜子面	柜子 1 顶面⑩⑪	形状	矩形
				高度	⑩同桌面 ⑪同地面高
				宽度	同桌面宽
				长度	50±20 cm
				方向	平行地面
			柜子 1 侧面⑫⑬⑭⑮	形状	矩形
				宽度	⑫⑭同桌面宽，⑬⑮同⑩长
				高度	同桌面高
				方向	垂直地面
				位置	⑫与⑤共面，⑬与④共面
		柜子 2	…	…	雷同柜子 1，略

本文算法中假定待重建物体都是基本几何体或者其表面只包含 3 个互相垂直方向上的矩形平面的非基本几何体（本文称这 3 个方向为物体的 3 个主方向）. 这主要是因为多数室内物体具有的这种特性，另外如果物体形状过于复杂，所需要的语义描述数据库会太大，也超过了本文拟研究的范围.

3.2　用户语音输入信息

高层重建需要用户对计算机输入一些对场景中待重建物体的高层感知信息，如物体名称、长、宽、高、颜色以及在场景中的大致方位等（这也是为什么要将点云变换到以地面为 $x-y$ 平面的坐标系，只有这样输入的物体的高度等信息才会有意义），这些信息是建立在用户对场景熟悉的基础上的. 用户输入的语音信息主要发挥两方面的作用：一是用于点云分割，将用户关注的物体从场景点云中分离出来；二是指导知识库完成物体的表面重建. 由于语音信息都要经过计算机理解才能转换成语义信息，故语音输入要求具有简略、充分、易于描述和理解的特点. 太多的输入信息会影响到算法的实用性；必须提供足够的信息以保证后续工作的顺利进行.

3.3　三维点云分割

要重建一个物体，首先要在底层重建的点云中分割出该物体的点云，这一点可以根据输入的语义信息辅助进行. 比如在某一个视角下，利用提供的物体的颜色、长、宽、高、在场景中的深度范围以及方位（前后、上下左右等）可以基本上分割出物体点云. 当然这些信息是由用户根据对场景的粗略了解估计而得到的，不可能准确分割. 但由于本文后续的算法采用直方图算法，少量的错误点并不会影响到表面的重建. 为了分割尽量准确，分割顺序需要遵循一定的原则：由近至远，由明显物体至不明显物体. 这是因为近处的物体深度较容易判断，从而为分割提供

比较可靠的信息;另外明显的物体也更容易与别的物体分开;远处的或特征不明显的物体可以根据与明显物体之间的方位关系等更多的信息进行分割.另一方面,物体是逐个分割重建的,即每分割出一个物体就将其重建,重建完成后,从原有点云中去除该物体的点云,以免影响后面物体的分割.如此往复,直至全部重建完毕.

3.4 根据语义重建表面

一旦点云分割出来,就能够利用知识库的信息对该物体的结构进行分析,进而重建表面.本文建立的知识库中的基元是二次曲面和平面.表面重建的过程就是根据输入语义有意识地按照知识库中的信息搜索对应于不同基元的点云,然后对搜索出的点云进行面拟合(曲面或平面),从而确定出基元的各项属性参数,以完成基元重建.基本几何体的基元重建相对比较简单,如篮球,其所对应基元是球面,直接用球面方程拟合篮球点云即可确定出篮球的半径和球心,从而完整地重建出场景中篮球表面.圆柱体对应的基元是 2 个平面(底面)和 1 个柱面(侧面),且底面和侧面相互垂直.在重建一个圆柱体时,只需要确定 2 个底面的位置并拟合出侧面截圆半径、圆心即可.长方体的基元是 6 个平面,分布在 3 个主方向上,需要根据已重建的空间信息(主要是直线段)确定出这 3 个方向以及分布在这 3 个主方向上的各个平面的属性参数.如同长方体一样,对于更为一般的室内物体,其表面也基本上由分布在 3 个主方向上的矩形平面片构成(如桌子).其重建过程类似长方体,包括主方向确定和沿着主方向确定平面(即基元)2 部分.

1) 确定主方向

首先地面的方向是 3 个主方向之一,本文用 N_0 表示该方向;然后根据采集平台的特点,只需确定另外 2 个主方向.通常,在已完成的底层重建结果中包含了部分线的信息,重建出的水平直线和垂直直线一般都分布在物体的边缘(即物体包含的平面的边缘),故可以直接用水平直线段来确定另外的 2 个主方向.具体做法是将$[0,\pi]$以 $\theta=5°$为步长分割成个 36 个 bins,累加按斜率属于不同 bins 的水平直线段的个数,选择含有水平直线段最多的 bin,并将其内的直线段的斜率进行平均,得到一个主方向 N_1.将 N_1 在$x-y$平面上旋转 90°,得到另一个主方向 N_2,将所有与某一主方向夹角小于 5°的水平直线段绕其中点旋转至对应的主方向上.

为了后续处理方便,本文将分割出的点云和直线段变换到分别以 N_0,N_1,N_2 为 z 轴、x 轴和 y 轴的坐标系内,主方向随之变为 3 个坐标轴方向.

2) 基元重建

基元重建即确定基元的所有属性(数值属性),当一个基元的所有属性都确定了,它就可以被重建出来.同样,当确定物体所有基元属性后,即可完整地重建该物体表面.本文通过逐个确定各基元在相应主方向上的位置来确定基元属性.

基元重建的顺序并不固定,但总的策略是先必要基元后可选基元.重建完必要基元,再根据点云以及可选基元与必要基元的关系(如位置关系、平行关系等)判断可选基元是否存在(点云数量大于某一阈值 τ,本文实验中该阈值取为 50),进一步决定该基元的各项属性.

与确定主方向的方法类似,本文也采用直方图算法沿着主方向确定各个基元位置.确定某一基元位置的具体步骤如下:首先由知识库得到其所在主方向,然后沿此主方向以 h 为步长将点云划分到若干 bins——B_i 中.根据

$$Cpbin = \{B_i \mid N(B_i) > \tau \ \&\ B_i \text{ 满足基元属性约束}\} \tag{1}$$

得到候选 bins 集合 $Cpbin$.最后利用

$$N(B_i) = \max_{j \in Cpbin} N(B_j), i \in Cpbin \tag{2}$$

在 $Cpbin$ 中选择含有点数最多的 B_i 作为基元所在 bin.将包含在 B_i 中的点在该主方向上的坐标值平均,得到该基元的位置,如图 5 所示.式(1)(2)中 $N(B_i)$是在 B_i 中点的数量.

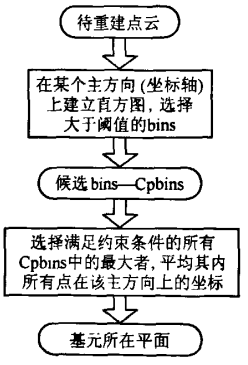


图 5 基元所在平面搜索流程

3.5 重建过程和实现

下面以如图 4 a L 形电脑桌为例来说明算法的具体实现过程.首先通过点线匹配重建得到了初步的场景点云和线段,并利用用户输入信息,通过 3.3 完成了点云分割.

按照电脑桌的知识信息，电脑桌有 3 个主方向. 通过第 3.4 节中的方法确定出 3 个主方向，将分割出的点云和直线段变换到分别以 N_0，N_1，N_2 为 z 轴、x 轴和 y 轴的坐标系内；然后在 3 个轴向上逐个搜索并重建基元. 为了统一处理，将直线段均匀离散成点集. 由于直线段一般是基元的边缘，所以其提供的结构信息较之离散的空间点云要更丰富；为了突出其重要性，在离散化的过程中步长应小一些.

由电脑桌知识，其所对应的基元包括桌面、侧面、柜子面和隔板. 本文的目标就是逐个将这些基元的位置等所有参数确定出来：

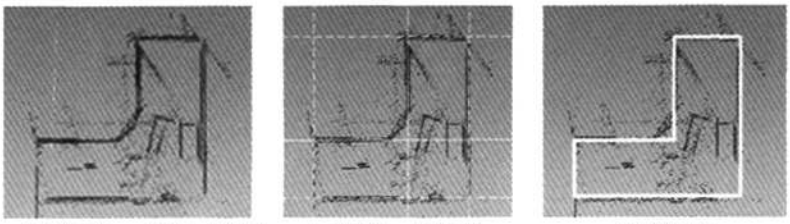
1) 在 z 方向上通过直方图算法确定桌面高度. 其中，候选集 $Cpbin$ 需要满足的属性约束是桌面的高度约束 $height\times(1-20\%)<i\times h<height\times(1+20\%)$；其中 i 是 B_i 的下标，h 是步长，$height$ 是用户输入的桌面高度(若没有用户输入，就用默认取值范围).

2) 沿着 x 方向和 y 方向确定出侧面. 将桌子点云分别在 x 轴方向和 y 轴方向上同样以 h 为步长划分成若干个 bins，找出候选 bins——$Cpbin$，这时 $Cpbin$ 需要满足电脑桌长度和宽度约束. 当然，此时并不知道哪个方向是长，哪个方向是宽. 通过判断满足不同约束的 bins 内含有点的数量确定长向和宽向：

$$length\times(1-20\%)<|i-j|\times h<length\times(1+20\%),$$
$$width\times(1-20\%)<|i-j|\times h<width\times(1+20\%).$$

其中，i，j 是 B_i，$B_j\in Cpbin$ 的下标；$width$，$length$ 分别输入的电脑桌的长、宽值. 因为每个方向都涉及到 2 个 bins，故通过选择包含点数之和最大的 2 个 bins 作为侧面所在 bins.

确定 4 个侧面位置的同时就找到了桌面的外轮廓，如图 6 b 中白色虚线条所示. 通过在 x，y 方向上选择到外侧边缘的距离满足桌面宽度约束的 2 个最大 bins，可以得到内侧边缘，如图 6 b 白色实线条所示. 内侧边缘与外侧边缘线相互之间相交即得到桌面边缘轮廓，如图 6 c 白色粗实线条所示. 这样，4 个侧面和桌面的参数就都确定了出来.



a 桌子点云在 x-y 平面上的投影　　b 桌面内、外轮廓线　　c 桌面轮廓线

图 6　桌面边缘的确定过程

3) 判断 2 个可选基元是否存在. 当找到桌面后，隔板的判断只需看侧面所在 bins 在桌面至隔板高度范围内是否有足够多的点(大于阈值 τ). 若有，则通过 z 方向直方图确定出隔板的高度(确定过程类似于桌面高度的确定，只是约束高度改为隔板的高度)；然后将侧面向上延伸至隔板高度. 因为按照电脑桌知识，隔板和对应的侧面共面.

由电脑桌知识，柜子在桌面以下紧靠在 2 个侧面内侧. 对于每一个柜子，其 2 个底面分别与桌面和地面共面，2 个侧面分别和电脑桌的 2 个侧面共面；另外还有一个与桌面内侧边缘共面，如图 4⑬所示. 于是，只需根据属性和点云信息判断出一个侧面，如图 4⑭所示的位置参数即可. 选择出相应部位的点云，判断其数量是否大于阈值 τ. 若是，则分别在 x，y 方向上利用直方图算法选择满足约束条件的含有点最多的 bin，并确定出该平面位置. 这里的约束条件是柜子的宽度.

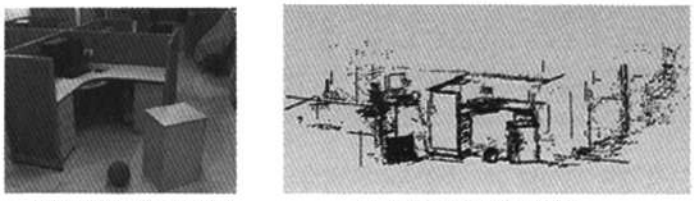
搜索完所有存在的基元后，根据确定的参数将其重建，即完成整个高层重建过程.

需要指出的是，无论是主方向的确定还是基元的确定，本文基本上都采用了直方图算法. 一方面该方法简单，整个运行过程基本上不花费什么时间；另一方面它具有比较好的鲁棒性，在存在少量的错误分割点或错误重建点的情况下也能保证正确地确定出各个基元. 在直方图中选择满足约束条件的最大者作为确定基元位置的 bin 原因如下：通常情况下，对于一个没有太多纹理的室内物体，其底层重建出的点和线段基本上分布在边缘附近，当然也都在物体的表面.

4 实验及结果

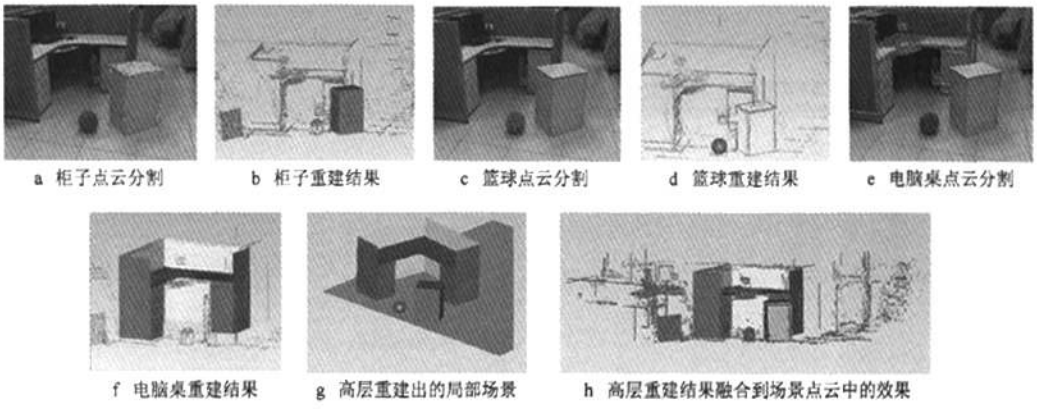
针对 2 个具体的室内场景局部，我们利用本文算法进行了重建，经过高层语义交互重建，原有的重建结果确实有了很大的改善.

实验 1. 图 7 a 所示为待重建场景的一个局部，本文利用文献[17]算法得到一些稀疏点云和水平、垂直直线段，如图 7 b 所示. 用户要在图 7 a 的视角下通过语音输入完成对篮球、柜子和一个有隔板的 L 形电脑桌的高层重建. 用户首先输入信息“场景最前面有一个灰白色柜子，高约 90 cm，宽约 50 cm，长约 50 cm”，计算机根据用户的提示搜索这一视角下深度最小的直线段和点云，并根据宽和长确定柜子所在的大致范围，结合颜色信息同时分割出这一范围内的点云和直线段，如图 8 a 所示；一旦柜子点云分割出来，利用水平直线段确定柜子的 2 个主方向，并根据柜子的知识，在 3 个主方向上确定出各个面的位置，从而确定出柜子各个面的数值属性，完成柜子表面的重建，如图 8 a 所示. 接着用户输入“柜子左侧有一个棕色篮球”，计算机根据颜色和方位、篮球的知识(半径)和已经重建出的柜子的深度信息，确定出篮球的点云，如图 8 c 所示；然后计算机通过判断篮球具有一个球面，进而利用球面拟合得到篮球的表面，如图 8 d 所示. 最后用户输入“篮球后面有一个 L 形电脑桌，在距其约 2 m 的范围内”，计算机根据输入的深度范围和电脑桌默认长宽等属性得到其所在大致的范围，从而分割出这部分点云和直线段，如图 8 e 所示. 进一步，利用第 3.5 节的方法重建出电脑桌的各个面，如图 8 f 所示. 因为电脑桌是标准的电脑桌，所以属性可以不输入. 表 3 所示为实验 1 部分位置的重建精度.



a 在用户当前视角下的场景　　b 底层重建的点云和直线段

图 7　实验 1 重建场景及底层重建结果



a 柜子点云分割　b 柜子重建结果　c 篮球点云分割　d 篮球重建结果　e 电脑桌点云分割

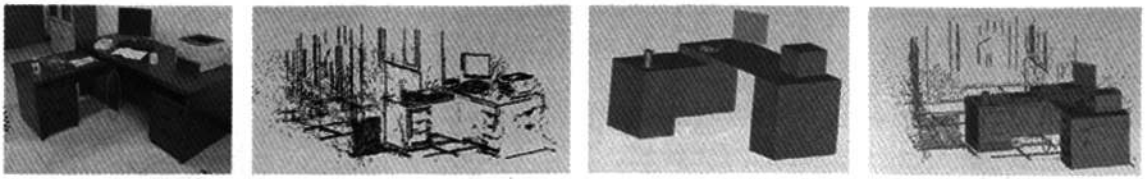
f 电脑桌重建结果　g 高层重建出的局部场景　h 高层重建结果融合到场景点云中的效果

图 8　实验 1 高层重建的过程及结果

表 3　实验 1 重建精度　　cm

测量位置	真值	测量值	绝对误差
桌面高度	77	78	1.0
桌面宽度	66	64.1584	1.8416
挡板高度	122	122.946	0.946
柜子宽度	48.2	49.6698	1.4698
柜子高度	70	74	4.0
篮球半径	12.3	12.17	0.13

实验 2. 图 9 所示为利用本文算法进行另一组实验的重建结果. 针对图 9 a 所示的场景，用户输入的信息依次为“场景最前面有一个直角办公桌”、“右侧桌面最前端上放着一个长方体打印机，高30 cm”、“右侧桌面后端放着一个 19 寸显示器”、“显示器左前方有一本字典，厚约 7 cm”、“左侧桌面最左端垂直放置一个圆形茶叶筒，高约 20 cm，直径约 10 cm”. 表 4 所示为部分位置的重建精度.



a 用户视角下的场景　b 底层重建的结果　c 高层重建结果　d 高层重建结果融合到场景点云中的效果

图 9　实验 2 的重建结果

表 4　实验 2 重建精度　　cm

测量位置	真值	测量值	绝对误差
打印机宽	35	33	2.0
打印机高	25	26.55	1.55
桌面宽度(左)	60	60.0031	0.0031
桌面宽度(右)	80	81.0006	1.0006
桌面高度(右)	76.5	74	2.5
显示器宽	41	39	2.0

从实验结果看到，通过语义交互对底层重建的结果确实有了很大程度上的改善，尤其是对于底层重建不可能重建出的遮挡部分，基于语义交互的重建也能根据知识库的信息予以弥补. 通过语义交互基本上能够得到比较完整的物体三维模型，事实上这一效果是目前所有不经过高层交互的视觉重建算法难以做到的. 当然，基于语义交互重建的结果也依赖于底层重建的结果，如果在底层重建时对于某一处表面没有得到任何信息或者信息不足，则语义交互重建也往往起不到改善效果，如图 9 a 中椭圆标注部分.

高层重建的精度依赖于底层标定、匹配、重建的精度，尤其是直线重建的精度. 直线上大量的点会直接影响到直方图中 bin 的判断，如实验 1 中柜子的高度精度较差，这主要是由柜子上边缘直线重建的误差造成的.

5　讨论及结论

在诸如室内场景三维重建过程中，由于环境缺乏纹理，单纯依靠纹理特征来进行重建往往只能得到非常稀疏的三维点云，很难满足实际应用的需求. 针对这种情况，本文利用底层重建的点云、直线段都基本分布在物体边缘，以及多数室内物体具有结构简单、固定的特点，尝试性地提出了一种基于语义交互的三维重建算法，即运用用户通过自然语言输入的关于场景的特定高层知识来辅助重建. 实验结果表明，这种思路具有一定的可行性，能够在一定程度上提高重建效果，包括对遮挡的处理. 在整个重建过程中，用户只需提供场景中物体的少量基本信息，其余的重建过程都可以自动完成.

需要特别强调的是，本文工作仅仅是一种尝试，是对语义交互下三维重建的一种探索，相关理论和算法都比较初步，并不完善，实验场景也相对简单. 但是，基于用户通过自然语言方式、通过用户对待重建场景特定高层语义的输入来进一步改善重建结果，具有交互自然方便、交互信息准确的优点，是一个有待进一步探索和深入研究的方向. 本文工作对从事三维重建的研究有一些参考作用.

参考文献(References)：

[1] Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis [C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2007: 1-8

[2] Furukawa Y, Curless B, Seitz S M, *et al*. Towards internet-scale multi-view stereo [C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2010: 1434-1441

[3] Furukawa Y, Curless B, Seitz S M, *et al*. Manhattan-world stereo [C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1422-1429

[4] Furukawa Y, Curless B, Seity S M, *et al*. Reconstruction building interiors from images [C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 80-87

[5] Heits G, Koller D. Learning spatial context: using stuff to find things [M] //Lecture Notes in Computer Science. Heidelberg: Springer, 2008, 5302: 30-43

[6] Desai C, Ramanan D, Fowlkes, C. Discriminative models for multi-class object layout [C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 229-236

[7] Torralba A, Murphy K P J, Freeman W T, *et al*. Context-based vision system for place and object recognition [C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003, 1: 273-280

[8] Singhal A, Luo J B, Zhu W Y. Probabilistic spatial context models for scene content understanding [C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2003, 1: 235-241

[9] Rabinovich A, Vedaldi A, Galleguillos C, *et al*. Objects in context [C] //Proceedings of the 11th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2007: 1-8

[10] Hoiem D, Efros A A, Hebert M. Putting objects in perspective [C] //Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2006: 2137-2144

[11] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene structure from a single still image [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840

[12] Boykov Y, Veksler O, Zabin R. Fast approximate energy minimization via graph cuts [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11): 1222-1239

[13] Woodford O, Torr P, Fitggibbon. Global stereo reconstruction under second-order smoothness priors [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2115-2128

[14] Nüchter A, Surmann H, Hertzberg J. Automatic model refinement for 3D reconstruction with mobile robots [C] // Proceedings of the 4th IEEE International Conference on 3-D Digital Imaging and Modeling. Los Alamitos: IEEE Computer Society Press, 2003: 394-401

[15] Cantzler H, Fisher R B, Devy M. Quality enhancement of reconstructed 3D models using coplanarity and constraints [M] // Lecture Notes in Computer Science. Heidelberg: Springer, 2002, 2449: 34-41

[16] Grau O. A scene analysis system for the generation of 3-D models [C] //Proceedings of International Conference on Recent Advance in 3-D Digital Imaging and Modeling. Los Alamitos: IEEE Computer Society Press, 1997: 221-228

[17] Zhang Feng, Shi Limin, Sun Fengmei, *et al*. An image based 3D reconstruction system for large indoor scenes [J]. Acta Automatica Sinica, 2010, 36(5): 625-633 (in Chinese)
(张 峰, 史利民, 孙凤梅, 等. 一种基于图像的室内大场景自动三维重建系统[J]. 自动化学报, 2010, 36(5): 625-633)

[18] Zhang F, Shi L M, Xu Z H, *et al*. A 3D reconstruction system of indoor scenes with rotating platform [C] // Proceedings of International Symposium on Computer Science and Computational Technology. Los Alamitos: IEEE Computer Society Press, 2008: 554-558