



(12) 发明专利申请

(10) 申请公布号 CN 112241411 A

(43) 申请公布日 2021.01.19

(21) 申请号 202011148183.3

(22) 申请日 2020.10.23

(71) 申请人 湖南省交通规划勘察设计院有限公司

地址 410200 湖南省长沙市望城区月亮岛路一段598号

(72) 发明人 贺耀北 刘婷婷 王永 杨云逸 李瑜 李文武

(74) 专利代理机构 湖南兆弘专利事务所(普通合伙) 43008

代理人 周长清

(51) Int. Cl.

G06F 16/22 (2019.01)

G06F 16/25 (2019.01)

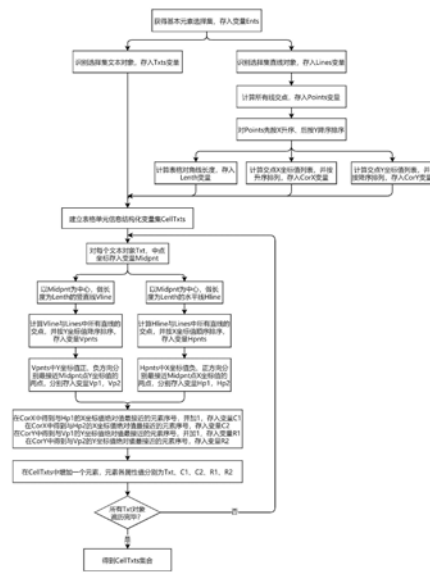
权利要求书2页 说明书5页 附图4页

(54) 发明名称

基于CAD基础元素的电子表格结构化识别与提取方法

(57) 摘要

本发明公开了一种基于CAD基础元素的电子表格结构化识别与提取方法,包括:S1:读入待输出图纸文件的文件数据;S2:用户框选构成表格形式的直线和文本对象,分别存为文本对象集合和直线对象集合;S3:对于选择几种所有的直线,两两之间计算交点,将交点存入交点集;进行排序;S4:计算交点集第一个元素与最后一个元素的距离,作为辅助线长度;S5:获得交点集所有元素的坐标值,按序排列;S6:对于文本集中的每个元素,进行循环遍历进行操作,获得对应的结构化信息;S7:完成所有文本集元素的结构化识别后,将所有的结构化单元信息数据提取到电子表格。本发明具有原理简单、易实现、处理效率高、适用范围广等优点。



CN 112241411 A

1. 一种基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,步骤包括:
步骤S1:打开AutoCAD读入待输出图纸文件的文件数据;
步骤S2:用户框选构成表格形式的直线和文本对象,分别存为文本对象集合和直线对象集合;
步骤S3:对于选择几种所有的直线,两两之间计算交点,将交点存入交点集;按照先按交点X坐标值,然后按交点Y坐标值对交点进行排序;
步骤S4:计算交点集第一个元素与最后一个元素的距离,作为辅助线长度;
步骤S5:获得交点集所有元素的X坐标值,存入X坐标集,并按升序排列;获得交点集所有元素的Y坐标值,存入Y坐标集,对Y坐标集按降序排列;
步骤S6:对于文本集中的每个元素,进行循环遍历进行操作,获得对应的结构化信息;
步骤S7:完成所有文本集元素的结构化识别后,将所有的结构化单元信息数据提取到电子表格。
2. 根据权利要求1所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S2中,若构成表格的元素中存在多段线、多行文字类型,先对所有对象执行分解命令,直至表格有直线和单行文本构成为止。
3. 根据权利要求2所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S2包括:
步骤S201:获取基本元素选择集,存入变量Ents;
步骤S202:识别选择集文本对象,存入Txts变量;识别选择集直线对象,存入Lines变量。
4. 根据权利要求3所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S3中包括:
步骤S301:计算所有线交点,存入Points变量;
步骤S302:对Points先按X升序、后按Y降序排序;即,计算表格对角线长度,存入Length变量;计算交点X坐标值表,并按升序排列,存入CorX变量;计算交点Y坐标值表,并按降序排列,存入CorY变量。
5. 根据权利要求1-4中任意一项所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S6中,获得对应的结构化信息的流程包括:
步骤S601:计算文本元素的中点坐标信息;
步骤S602:以中点坐标为中心点,按辅助线长度,作一根竖直辅助线;
步骤S603:计算该竖直辅助线与直线对象集合所有元素的交点,并记录具有与中点Y坐标值正负方向最接近Y坐标值的两个交点;
步骤S604:在Y坐标集中得到与上述两交点Y坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始列编号,较大的序号作为该文本所占单元格的终止列编号;
步骤S605:以中点坐标为中心点,按辅助线长度,作一根水平辅助线;
步骤S606:计算该水平辅助线与直线对象集合所有元素的交点,并记录具有与中点X坐标值正负方向最接近X坐标值的两个交点;
步骤S607:在X坐标集中得到与上述两交点X坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始行编号,较大的序号作为该文本所占单元格的终止行编号;

步骤S608:文本元素的文本内容,以及起点起始行编号、终止行编号、起始列编号、终止列编号,构成了一个结构化单元信息数据。

6. 根据权利要求5所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S601中对每个文本对象Txt,中点坐标存入变量Mdipt;所述步骤S602中以Mdipt为中心,做长度为Length的竖直线Vline;所述步骤S603中计算Vline与Lines中所有直线的交点,并按Y坐标值降序排序,存入变量Vpnts;所述步骤S604中Vpnts中Y坐标值正、负方向分别最接近Mdipt点Y坐标值的两点,分别存入变量Vp1、Vp2;所述步骤S605中以Mdipt为中心,做长度为Length的水平线Hline;所述步骤S606中计算Hline与Lines中所有直线的交点,并按X坐标值升序排序,存入变量Hpnts;所述步骤S607中Vpnts中X坐标值正、负方向分别最接近Mdipt点X坐标值的两点,分别存入变量Hp1、Hp2。

7. 根据权利要求6所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S608中包括:

在CorX中得到与Hp1的X坐标值绝对值最接近的元素序号,并加1,存入变量C1;
在CorX中得到与Hp2的X坐标值绝对值最接近的元素序号,并加1,存入变量C2;
在CorY中得到与Vp1的Y坐标值绝对值最接近的元素序号,并加1,存入变量R1;
在CorY中得到与Vp2的Y坐标值绝对值最接近的元素序号,并加1,存入变量R2;
在CellTxts中增加一个元素,元素各属性值分别为Txts、C1、C2、R1、R2。

8. 根据权利要求1-4中任意一项所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S7中对于结构化单元信息数据的处理包括:

a) 将EXCEL单元格中对应结构化单元信息数据的起始行编号、终止行编号、起始列编号、终止列编号的单元格进行合并;

b) 在所有结构化单元信息数据中得到与当前结构化单元信息数据起始行编号、终止行编号、起始列编号、终止列编号相同的元素,并按文本中心点Y坐标值降序排列,这些元素将被认定为属于同一单元格的不同行文本,按顺序加入单元格文本;每加入一个元素,在单元格文本末尾添加一个换行符号。

9. 根据权利要求8所述的基于CAD基础元素的电子表格结构化识别与提取方法,其特征在于,所述步骤S7的流程为:

步骤S701:启动EXCEL接口,提取CellTxts集合;

步骤S702:对每个CellTxt对象,先把EXCEL单元格(R1、C1)至(R2、C2)合并为一个单元格,并设为当前单元格Cell;然后,在CellTxts中得到与当前CellTxt对象C1、C2、R1、R2均值相同的集合,存入变量Temp;接下来对Temp中所有对象按Txt重点Y值降序排列,依次将Temp集合每个CellTxt对象的文本加入到字符串Str,并添加换行符号;最后在EXCEL单元格Cell中写入字符串Str;

步骤S703:所有CellTxt对象遍历完毕,完成执行。

基于CAD基础元素的电子表格结构化识别与提取方法

技术领域

[0001] 本发明主要涉及到计算机辅助设计技术领域,特指一种基于CAD基础元素的表格结构化识别与提取方法。

背景技术

[0002] CAD (Computer Aided Design) 计算机辅助设计,是计算机技术的一个重要的应用领域。AutoCAD是美国Autodesk公司开发的交互式绘图软件,用于二维及三维设计、绘图的系统工具,用户可以使用它来创建、浏览、管理、打印、输出、共享富含信息的设计图形。作为通用型的制图软件,AutoCAD广泛用于各个行业的设计工作。

[0003] AutoCAD图纸设计信息主要分为图形和表格两大类。其中,表格主要承载各类工程数量信息,是设计表达的主要内容,对于材料预备、造价控制、进度控制等工程管理各方面具有重要作用。由于设计人员的习惯和技术资料积累,在工程实践中有大量的表格是以基础元素直线构成的表格线和基础元素单行文本构成的表格内容这类形式存在。这类由基础元素构成的表格,具有表格形式的外观,实际上却是直线、单行或多行文本的松散集合,没有结构化数据,也无法与电子表格程序,如EXCEL等进行交互,制约了提高设计生产效率的提高。

[0004] 有从业者也提出过尝试采用程序对基于CAD基础元素的表格结构化识别的方法,但普遍存在算法复杂、限制条件多、识别准确度低的问题。

发明内容

[0005] 本发明要解决的技术问题就在于:针对现有技术存在的技术问题,本发明提供一种原理简单、易实现、处理效率高、适用范围广的基于CAD基础元素的电子表格结构化识别与提取方法。

[0006] 为解决上述技术问题,本发明采用以下技术方案:

[0007] 一种基于CAD基础元素的电子表格结构化识别与提取方法,其步骤包括:

[0008] 步骤S1:打开AutoCAD读入待输出图纸文件的文件数据;

[0009] 步骤S2:用户框选构成表格形式的直线和文本对象,分别存为文本对象集合和直线对象集合;

[0010] 步骤S3:对于选择几种所有的直线,两两之间计算交点,将交点存入交点集;按照先按交点X坐标值,然后按交点Y坐标值对交点进行排序;

[0011] 步骤S4:计算交点集第一个元素与最后一个元素的距离,作为辅助线长度;

[0012] 步骤S5:获得交点集所有元素的X坐标值,存入X坐标集,并按升序排列;获得交点集所有元素的Y坐标值,存入Y坐标集,对Y坐标集按降序排列;

[0013] 步骤S6:对于文本集中的每个元素,进行循环遍历进行操作,获得对应的结构化信息;

[0014] 步骤S7:完成所有文本集元素的结构化识别后,将所有的结构化单元信息数据提

取到电子表格。

[0015] 作为本发明方法的进一步改进:所述步骤S2中,若构成表格的元素中存在多段线、多行文字类型,先对所有对象执行分解命令,直至表格有直线和单行文本构成为止。

[0016] 作为本发明方法的进一步改进:所述步骤S2包括:

[0017] 步骤S201:获取基本元素选择集,存入变量Ents;

[0018] 步骤S202:识别选择集文本对象,存入Txts变量;识别选择集直线对象,存入Lines变量。

[0019] 作为本发明方法的进一步改进:所述步骤S3中包括:

[0020] 步骤S301:计算所有线交点,存入Points变量;

[0021] 步骤S302:对Points先按X升序、后按Y降序排序;即,计算表格对角线长度,存入Length变量;计算交点X坐标值表,并按升序排列,存入CorX变量;计算交点Y坐标值表,并按降序排列,存入CorY变量。

[0022] 作为本发明方法的进一步改进:所述步骤S6中,获得对应的结构化信息的流程包括:

[0023] 步骤S601:计算文本元素的中点坐标信息;

[0024] 步骤S602:以中点坐标为中心点,按辅助线长度,作一根竖直辅助线;

[0025] 步骤S603:计算该竖直辅助线与直线对象集合所有元素的交点,并记录具有与中点Y坐标值正负方向最接近Y坐标值的两个交点;

[0026] 步骤S604:在Y坐标集中得到与上述两交点Y坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始列编号,较大的序号作为该文本所占单元格的终止列编号;

[0027] 步骤S605:以中点坐标为中心点,按辅助线长度,作一根水平辅助线;

[0028] 步骤S606:计算该水平辅助线与直线对象集合所有元素的交点,并记录具有与中点X坐标值正负方向最接近X坐标值的两个交点;

[0029] 步骤S607:在X坐标集中得到与上述两交点X坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始行编号,较大的序号作为该文本所占单元格的终止行编号;

[0030] 步骤S608:文本元素的文本内容,以及起点起始行编号、终止行编号、起始列编号、终止列编号,构成了一个结构化单元信息数据。

[0031] 作为本发明方法的进一步改进:所述步骤S601中对每个文本对象Txt,中点坐标存入变量Mdipnt;所述步骤S602中以Mdipnt为中心,做长度为Length的竖直线Vline;所述步骤S603中计算Vline与Lines中所有直线的交点,并按Y坐标值降序排序,存入变量Vpnts;所述步骤S604中Vpnts中Y坐标值正、负方向分别最接近Mdipnt点Y坐标值的两点,分别存入变量Vp1、Vp2;所述步骤S605中以Mdipnt为中心,做长度为Length的水平线Hline;所述步骤S606中计算Hline与Lines中所有直线的交点,并按X坐标值升序排序,存入变量Hpnts;所述步骤S607中Vpnts中X坐标值正、负方向分别最接近Mdipnt点X坐标值的两点,分别存入变量Hp1、Hp2。

[0032] 作为本发明方法的进一步改进:所述步骤S608中包括:

[0033] 在CorX中得到与Hp1的X坐标值绝对值最接近的元素序号,并加1,存入变量C1;

- [0034] 在CorX中得到与Hp2的X坐标值绝对值最接近的元素序号,并加1,存入变量C2;
- [0035] 在CorY中得到与Vp1的Y坐标值绝对值最接近的元素序号,并加1,存入变量R1;
- [0036] 在CorY中得到与Vp2的Y坐标值绝对值最接近的元素序号,并加1,存入变量R2;
- [0037] 在CellTxts中增加一个元素,元素各属性值分别为Txts、C1、C2、R1、R2。
- [0038] 作为本发明方法的进一步改进:所述步骤S7中对于结构化单元信息数据的处理包括:
- [0039] a) 将EXCEL单元格中对应结构化单元信息数据的起始行编号、终止行编号、起始列编号、终止列编号的单元格进行合并;
- [0040] b) 在所有结构化单元信息数据中得到与当前结构化单元信息数据起始行编号、终止行编号、起始列编号、终止列编号相同的元素,并按文本中心点Y坐标值降序排列,这些元素将被认定为属于同一单元格的不同行文本,按顺序加入单元格文本;每加入一个元素,在单元格文本末尾添加一个换行符号。
- [0041] 作为本发明方法的进一步改进:所述步骤S7的流程为:
- [0042] 步骤S701:启动EXCEL接口,提取CellTxts集合;
- [0043] 步骤S702:对每个CellTxt对象,先把EXCEL单元格(R1、C1)至(R2、C2)合并为一个单元格,并设为当前单元格Cell;然后,在CellTxts得到与当前CellTxt对象C1、C2、R1、R2均值相同的集合,存入变量Temp;接下来对Temp中所有对象按Txt重点Y值降序排列,依次将Temp集合每个CellTxt对象的文本加入到字符串Str,并添加换行符号;最后在EXCEL单元格Cell中写入字符串Str;
- [0044] 步骤S703:所有CellTxt对象遍历完毕,完成执行。
- [0045] 与现有技术相比,本发明的优点在于:
- [0046] 本发明的基于CAD基础元素的电子表格结构化识别与提取方法,原理简单、易实现、处理效率高、适用范围广,其针对由CAD基础元素直线、单行文本构成的表格,可以简单的运算即可以得到结构化单元信息数据,具有运算效率高、识别准确度高的优点。由于所有形式的CAD表格最后都可以分解为由直线、单行文本构成的表格,故本发明方法可以实现对所有形式表格的快速识别。

附图说明

- [0047] 图1是本发明在具体应用实施例中的流程示意图。
- [0048] 图2是本发明在具体应用实施例中表格提取实施的流程示意图。
- [0049] 图3是本发明在具体应用实施例中CAD基础元素的表格界面示意图。
- [0050] 图4是本发明在具体应用实施例中提取得到电子表格界面示意图。

具体实施方式

- [0051] 以下将结合说明书附图和具体实施例对本发明做进一步详细说明。
- [0052] 如图1所示,本发明的基于CAD基础元素的电子表格结构化识别与提取方法,其步骤包括:
- [0053] 步骤S1:打开AutoCAD读入其支持格式(如DWG、DXF等)的待输出图纸文件的文件数据。

[0054] 步骤S2:用户框选构成表格形式的直线和文本对象,分别存为文本对象集合和直线对象集合;在具体应用时,若构成表格的元素中存在多段线、多行文字类型,可以先对所有对象执行分解命令,直至表格有直线和单行文本构成为止。

[0055] 步骤S3:对于选择几种所有的直线,两两之间计算交点,将交点存入交点集,按照先按交点X坐标值,然后按交点Y坐标值对交点进行排序。

[0056] 步骤S4:计算交点集第一个元素与最后一个元素的距离,作为辅助线长度。

[0057] 步骤S5:获得交点集所有元素的X坐标值,存入X坐标集,并按升序排列;获得交点集所有元素的Y坐标值,存入Y坐标集,由于AutoCAD图纸Y正坐标向下,故对Y坐标集按降序排列。

[0058] 步骤S6:对于文本集中的每个元素,进行循环遍历进行操作,获得对应的结构化信息,即建立表格单元信息结构化变量集CellTxts。

[0059] 步骤S7:完成所有文本集元素的结构化识别后,可以将所有的结构化单元信息数据提取到EXCEL电子表格。

[0060] 在具体应用实例中,所述步骤S2中包括:

[0061] 步骤S201:获取基本元素选择集,存入变量Ents;

[0062] 步骤S202:识别选择集文本对象,存入Txts变量;识别选择集直线对象,存入Lines变量。

[0063] 在具体应用实例中,所述步骤S3中包括:

[0064] 步骤S301:计算所有线交点,存入Points变量;

[0065] 步骤S302:对Points先按X升序、后按Y降序排序;即,计算表格对角线长度,存入Length变量;计算交点X坐标值表,并按升序排列,存入CorX变量;计算交点Y坐标值表,并按降序排列,存入CorY变量。

[0066] 在具体应用实例中,所述步骤S6中,获得对应的结构化信息的流程包括:

[0067] 步骤S601:计算文本元素的中点坐标信息;即,对每个文本对象Txt,中点坐标存入变量Mdipnt;

[0068] 步骤S602:以中点坐标为中心点,按辅助线长度,作一根竖直辅助线;即,以Mdipnt为中心,做长度为Length的竖直线Vline;

[0069] 步骤S603:计算该竖直辅助线与直线对象集合所有元素的交点,并记录具有与中点Y坐标值正负方向最接近Y坐标值的两个交点;即,计算Vline与Lines中所有直线的交点,并按Y坐标值降序排序,存入变量Vpnts;

[0070] 步骤S604:在Y坐标集中得到与上述两交点Y坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始列编号,较大的序号作为该文本所占单元格的终止列编号;即,Vpnts中Y坐标值正、负方向分别最接近Mdipnt点Y坐标值的两点,分别存入变量Vp1、Vp2;

[0071] 步骤S605:以中点坐标为中心点,按辅助线长度,作一根水平辅助线;以Mdipnt为中心,做长度为Length的水平线Hline;

[0072] 步骤S606:计算该水平辅助线与直线对象集合所有元素的交点,并记录具有与中点X坐标值正负方向最接近X坐标值的两个交点;计算Hline与Lines中所有直线的交点,并按X坐标值升序排序,存入变量Hpnts;

[0073] 步骤S607:在X坐标集中得到与上述两交点X坐标值相同的元素序号,较小的序号需加1作为该文本所占单元格的起始行编号,较大的序号作为该文本所占单元格的终止行编号;Vpnts中X坐标值正、负方向分别最接近Mdipnt点X坐标值的两点,分别存入变量Hp1、Hp2;

[0074] 步骤S608:文本元素的文本内容,以及起点起始行编号、终止行编号、起始列编号、终止列编号,构成了一个结构化单元信息数据。即:

[0075] 在CorX中得到与Hp1的X坐标值绝对值最接近的元素序号,并加1,存入变量C1;

[0076] 在CorX中得到与Hp2的X坐标值绝对值最接近的元素序号,并加1,存入变量C2;

[0077] 在CorY中得到与Vp1的Y坐标值绝对值最接近的元素序号,并加1,存入变量R1;

[0078] 在CorY中得到与Vp2的Y坐标值绝对值最接近的元素序号,并加1,存入变量R2;

[0079] 在CellTxts中增加一个元素,元素各属性值分别为Txts、C1、C2、R1、R2。

[0080] 在具体应用实例中,所述步骤S7中,对于结构化单元信息数据的处理包括:

[0081] a) 将EXCEL单元格中对应结构化单元信息数据的起始行编号、终止行编号、起始列编号、终止列编号的单元格进行合并;

[0082] b) 在所有结构化单元信息数据中得到与当前结构化单元信息数据起始行编号、终止行编号、起始列编号、终止列编号相同的元素,并按文本中心点Y坐标值降序排列,这些元素将被认定为属于同一单元格的不同行文本,按顺序加入单元格文本;每加入一个元素,在单元格文本末尾添加一个换行符号。

[0083] 在一个实际应用实例中,所述步骤S7的流程为:

[0084] 步骤S701:启动EXCEL接口,提取CellTxts集合;

[0085] 步骤S702:对每个CellTxt对象,先把EXCEL单元格(R1、C1)至(R2、C2)合并为一个单元格,并设为当前单元格Cell;然后,在CellTxts得到与当前CellTxt对象C1、C2、R1、R2均值相同的集合,存入变量Temp;接下来对Temp中所有对象按Txt重点Y值降序排列,依次将Temp集合每个CellTxt对象的文本加入到字符串Str,并添加换行符号;最后在EXCEL单元格Cell中写入字符串Str;

[0086] 步骤S703:所有CellTxt对象遍历完毕,完成执行。

[0087] 生成的实例样本,如图3的所示的CAD基础元素的表格界面示意图,最终生成如图4所示的提取得到电子表格界面示意图。

[0088] 以上仅是本发明的优选实施方式,本发明的保护范围并不仅局限于上述实施例,凡属于本发明思路下的技术方案均属于本发明的保护范围。应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理前提下的若干改进和润饰,应视为本发明的保护范围。

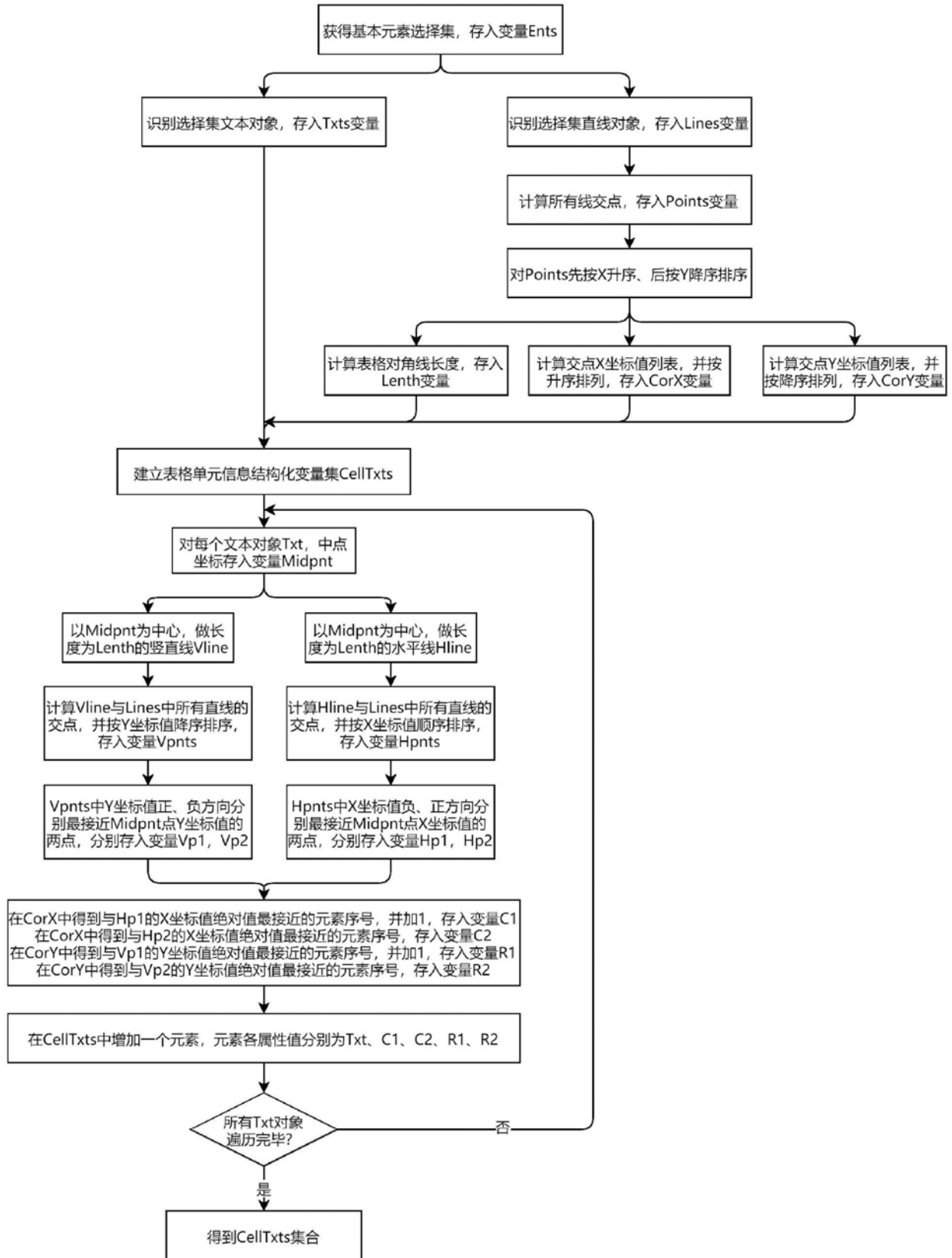


图1

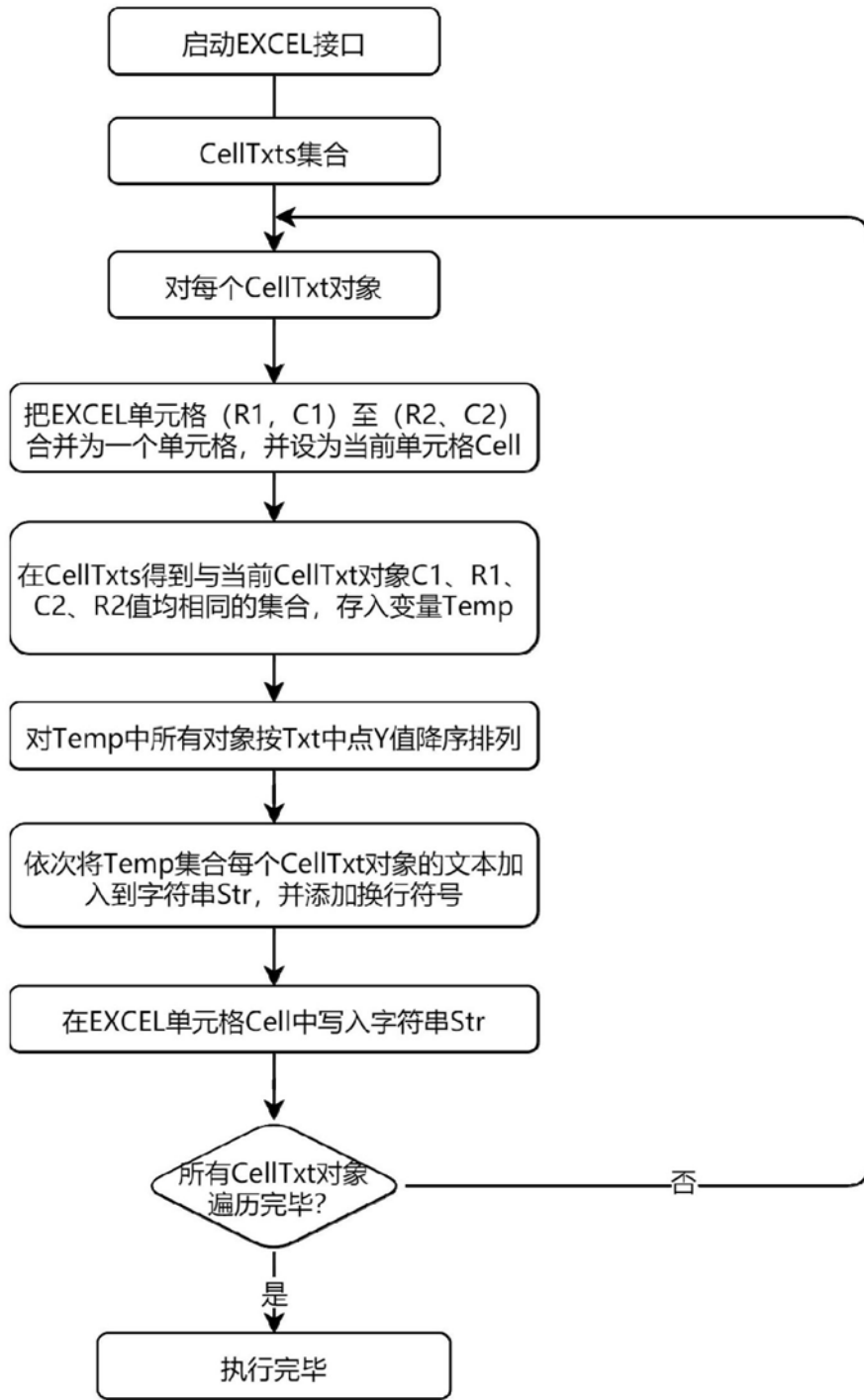


图2

名称	编号	规格 (mm)	数量	单件重 (kg)	共重 (kg)
腹板	N1	□3461×14×12486	1	4469	4469
	N2	□3346×16×6828	2	4276	8552
水平加劲	N3	□140×16×12088	1	213	213
水平加劲	N4	□140×10×12086	1	133	133
竖向加劲	N5	□120×10×2787	8	26	210
水平加劲	N6	□140×10×6000	2	66	132
水平加劲	N7	□140×10×3000	4	33	132
水平加劲	N8	□140×10×3200	4	35	141
竖向加劲	N9	□120×10×1805	12	17	204
人孔加劲	N10	□120×10×1805	4	17	68
电缆孔加劲	N11	□120×10×4104	2	39	77
横梁底板	N12	□600×25×12500	1	1472	1472
角点加劲					178
合计					15980
焊缝 (1.5%)					240
总计					16220

图3

	A	B	C	D	E	F
1	名称	编号	规格 (mm)	数量	单件重 (kg)	共重 (kg)
2	腹板	N1	□3461×14×12486	1	4469	4469
3		N2	□3346×16×6828	2	4276	8552
4	水平加劲	N3	□140×16×12088	1	213	213
5	水平加劲	N4	□140×10×12086	1	133	133
6	竖向加劲	N5	□120×10×2787	8	26	210
7	水平加劲	N6	□140×10×6000	2	66	132
8	水平加劲	N7	□140×10×3000	4	33	132
9	水平加劲	N8	□140×10×3200	4	35	141
10	竖向加劲	N9	□120×10×1805	12	17	204
11	人孔加劲	N10	□120×10×1805	4	17	68
12	电缆孔加劲	N11	□120×10×4104	2	39	77
13	横梁底板	N12	□600×25×12500	1	1472	1472
14	角点加劲					178
15	合计					15980
16	焊缝 (1.5%)					240
17	总计					16220

图4