



(21)申请号 201810070349.0

(22)申请日 2018.01.24

(71)申请人 陕西科技大学

地址 710021 陕西省西安市未央区大学园
区陕西科技大学

(72)发明人 孙连山 侯涛 张沙沙

(74)专利代理机构 西安智大知识产权代理事务
所 61215

代理人 段俊涛

(51)Int.Cl.

G06K 9/00(2006.01)

G06K 9/46(2006.01)

G06K 9/62(2006.01)

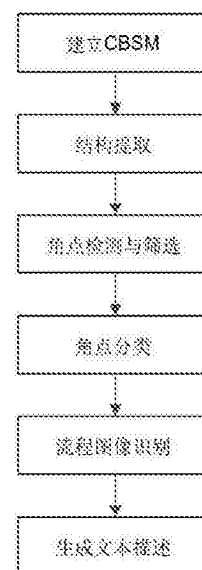
权利要求书4页 说明书7页 附图3页

(54)发明名称

一种流程图像识别方法

(57)摘要

本发明提出了一种流程图像识别方法,其技术特征在于:首先建立一个基于角点的流程图结构模型(CBSM),形式地定义用于描述流程图结构元素的各种角点类型、组合规则及约束;然后进行文图分割提取流程图结构,检测流程图结构的角点并提取角点邻域的网格特征和外围特征,采用SVM分类器实现角点分类;基于CBSM识别流程图结构;最后将流程图像识别结果以文本形式输出。本发明基于角点特征识别流程图结构元素,能够解决现有流程图像识别技术未能处理好的文图粘连及断边问题,为流程图像识别提供新方法。



1. 一种流程图像识别方法,其特征在于,包括如下步骤:

S1建立一个基于角点的流程图结构模型(CBSM),定义描述流程图结构元素的角点类型、角点组合规则以及角点组合约束;

S2提取流程图结构;

S3对流程图结构进行角点检测和筛选;

S4得到步骤S3中角点的邻域特征并训练角点分类器,对角点进行分类;

S5基于CBSM对流程图结构元素进行识别;

S6文本识别以及结构描述生成。

2. 根据权利要求1所述流程图像识别方法,其特征在于,所述步骤S1中建立CBSM,定义描述流程图结构元素的角点类型、角点组合规则以及角点组合约束包括:

S11将描述流程图结构元素的角点分为独立型和连接型两大类,命名各角点类型并建立流程图像角点分类表;

S12建立角点组合规则;

S13建立角点组合约束。

3. 根据权利要求2所述流程图像识别方法,其特征在于,所述步骤S11中,经典图元包括矩形、菱形和椭圆形,定义矩形为R图元,菱形为D图元,椭圆形为E图元,则描述经典图元的角点及其命名规则如下:

将仅属于单个图元或连接线的角点定义为独立型角点,其中对R图元的四个角点和折线的角点,分别命名为:

\top :R-1、 \sqcap :R-2、 \bot :R-3、 \sqcup :R-4;

将位于图元与连接线或连接线与连接线相接处的角点定义为连接型角点,其中对R图元和连接线相接处以及连接线与连接线相接处的角点,分别命名为:

\perp :Rb-1、 \top :Rb-2、 \sqcap :Rb-3、 \sqcup :Rb-4;

将E图元的独立型角点根据位置分别命名为:

\curvearrowright :E1-1、 \curvearrowleft :Er-1、 \curvearrowright :E1-2、 \curvearrowleft :Er-2;

将组成E图元的连接型角点分别命名为:

\perp :Rb-1、 \top :Rb-2;

将D图元的独立型角点按照上下左右的顺序依次命名为D-1、D-2、D-3、D-4;

将D图元的连接型角点按照上下左右的顺序依次命名为Db-1、Db-2、Db-3、Db-4;

所述步骤S12中,根据图元几何关系和角点相对位置关系,定义组合规则如下:

角点相对位置关系:以 $C_i, C_j (i \neq j)$ 表示两个不同角点, $x(C_i)$ 表示角点 C_i 横坐标, $y(C_i)$ 表示角点 C_i 纵坐标,角点 C_j 相对于角点 C_i 的位置关系有8种,分别为左上、左、左下、下、右下、右、右上、上,依次分别表示为u1、l、d1、d、dr、r、ur、u,每种位置关系对应一个判定表达式;

以 $r(C_i, C_j)$ 表示 C_j 在 C_i 的右方位置,满足 $x(C_i) < x(C_j)$; $d(C_i, C_j)$ 表示 C_j 在 C_i 下方位置,满足 $y(C_i) < y(C_j)$; $d1(C_i, C_j)$ 表示 C_j 在 C_i 的左下方位置,满足 $(x(C_i) > x(C_j)) \wedge (y(C_i) < y(C_j))$; 其中, $r(C_i, C_j)$ 等价于 $l(C_j, C_i)$, $d(C_i, C_j)$ 等价于 $u(C_j, C_i)$, $dr(C_i, C_j)$ 等价于 $u1(C_j, C_i)$, $d1(C_i, C_j)$ 等价于 $ur(C_j, C_i)$, $d(C_i, C_j)$ 等价于 $u(C_j, C_i)$, $r(C_i, C_j)$ 等价于 $l(C_j, C_i)$;

在流程图结构元素中,角点间几何关系描述为水平关系H或者垂直关系V, $H(C_i, C_j)$ 表示 C_i 与 C_j 处于同一条水平线上,满足 $y(C_i) = y(C_j)$; $V(C_i, C_j)$ 表示 C_i 与 C_j 处于同一条垂直线上,

满足 $x(C_i) = y(C_j)$; 其中, $H(C_i, C_j)$ 等价于 $H(C_j, C_i)$, $V(C_i, C_j)$ 等价于 $V(C_j, C_i)$;

R图元的角点组合为 $R-1 \wedge R-2 \wedge R-3 \wedge R-4$;

R图元的角点相对位置关系表示如下:

$r(R-1, R-2) \wedge d(R-1, R-3) \wedge dr(R-1, R-4) \wedge dl(R-2, R-3) \wedge d(R-2, R-4) \wedge r(R-3, R-4)$;

R图元的角点间几何关系表示如下:

$H(R-1, R-2) \wedge V(R-1, R-3) \wedge H(R-3, R-4) \wedge V(R-2, R-4)$;

E图元的角点组合表示为 $E1-1 \wedge E1-2 \wedge Er-1 \wedge Er-2$;

E图元的角点之间相对位置表示如下:

$r(E1-1, Er-1) \wedge d(E1-1, E1-2) \wedge dr(E1-1, Er-2) \wedge dl(Er-1, E1-2) \wedge d(Er-1, Er-2) \wedge r(E1-2, Er-2)$;

E图元的角点间几何关系表示如下:

$H(E1-1, Er-1) \wedge V(E1-1, E1-2) \wedge H(E1-2, Er-2) \wedge V(Er-1, Er-2)$;

D图元的角点组合表示为 $D*-1 \wedge D*-2 \wedge D*-3 \wedge D*-4$, D^* -表示D-型角点或Db-型角点, D-型角点为独立型角点, Db-型角点为连接型角点;

D图元的角点相对位置关系表示如下:

$d(D*-1, D*-2) \wedge dl(D*-1, D*-3) \wedge dr(D*-1, D*-4) \wedge ul(D*-2, D*-3) \wedge ur(D*-2, D*-4) \wedge r(D*-3, D*-4)$;

D图元的角点间几何关系表示如下: $H(D*-3, D*-4) \wedge V(D*-1, D*-2)$;

所述步骤S13中, 定义描述流程图结构元素的角点组合所满足的约束, 具体包括角点类型和数量约束、角点间的几何约束以及由角点组合所确定的图元的宽高约束, 分别定义如下:

各图元的角点组合中角点类型和数量的约束如下:

$$\begin{array}{l} \text{R图元} \left\{ \begin{array}{l} \text{独立型角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 2; \\ \text{同一类型角点数量} < 2; \end{array} \right. \quad \text{E图元} \left\{ \begin{array}{l} \text{独立型角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 1; \\ \text{同一类型角点数量} < 2; \end{array} \right. \\ \text{D图元} \left\{ \begin{array}{l} \text{角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 2; \\ \text{同一类型角点数量} < 2; \end{array} \right. \end{array}$$

角点组合中的角点间几何约束指的是两个角点在一定像素范围内呈水平关系H或者垂直关系V, 分别定义为: $H(C_i, C_j) :: |y(C_i) - y(C_j)| < d$, $V(C_i, C_j) :: |x(C_i) - x(C_j)| < d$, d 为修正在角点检测过程中出现的位置偏差所设置的松弛因子, 经验取值为5px;

角点组合所构成图元的宽和高约束如下所述, 若流程图像的整体宽和高分别为 W, H , R_{width} 表示R图元的宽, R_{height} 表示R图元的高, 则:

$$\text{R图元: } R_{width} < W \wedge R_{height} < \frac{H}{2};$$

E_{width} 表示E图元的宽, E_{height} 表示E图元的高, 则:

$$\text{E图元: } E_{width} < \frac{W}{2} \wedge E_{height} < \frac{H}{3};$$

D_{width} 表示D图元的宽, D_{height} 表示D图元的高,则:

$$D\text{图元}: E_{width} < W \wedge E_{height} < \frac{H}{2}。$$

4. 根据权利要求1所述流程图像识别方法,其特征在于,所述步骤S2中提取流程图结构的方法:

对原始流程图像进行二值化、降噪处理,通过连通域提取方法进行图文分割得到原始流程图像中结构层,对提取的流程图结构进行单像素化处理。

5. 根据权利要求1所述流程图像识别方法,其特征在于,所述步骤S3中对流程图结构进行角点检测和筛选的步骤为:

S31采用曲率尺度空间(Curvature Scale Space,CSS)算法利用曲率检测出与直线型元素相关的角点,过滤掉曲线型结构元素上的圆角点和虚假角点;

S32对采用CSS算法检测得到的角点进行边界搜索并剪裁掉边界包围的区域,得到曲线型结构元素所在的区域,然后采用Harris角点检测算法对该区域进行再次检测,得到曲线型结构元素上的角点;

S33在检测过程中,曲线上容易产生角点聚簇现象,采用基于欧式距离的冗余角点筛选方法,计算所有点之间的欧式距离,当角点间的距离小于指定阈值时,删掉其中之一以减少冗余。

6. 根据权利要求5所述流程图像识别方法,其特征在于,所述步骤S31中,直线型元素包括矩形、菱形和直线连接线,其边缘轮廓局部曲率变化明显;曲线型元素多指圆、椭圆和曲线连接线,其边缘轮廓局部曲率变化不明显;采用CSS算法在同一尺度不同曲率下对流程图结构进行角点检测,可以得到直线型结构元素上的绝大多数角点、曲线型结构元素上的部分角点以及部分虚假角点,将曲线型结构元素上的角点视为圆角点并同虚假角点一起过滤掉,圆角点及虚假角点的判定方法如下:

$$R_c = \begin{cases} 1, & K(u) \leq T(u) \\ 0, & \text{other} \end{cases}, \text{其中} u \text{为候选角点的位置参数}, K(u) \text{是候选角点的曲率}, T(u)$$

为与角点支持域自适应的动态局部阈值,与候选角点处的局部平均曲率成正比,当 $R_c = 1$ 时表示角点为圆角点,给予滤除;

$$C_c \text{ is } \begin{cases} \text{true corner}, & \angle C_c \leq \theta_{obtuse} \\ \text{false corner}, & \angle C_c > \theta_{obtuse} \end{cases}, \text{其中} C_c \text{为需要判定的候选角点}, \angle C_c \text{为角}$$

点 C_c 的角, θ_{obtuse} 为真角点的最大钝角值, θ_{obtuse} 的经验取值为 162° ,当 $C_c > \theta_{obtuse}$ 时, C_c 为虚假角点。

7. 根据权利要求1所述流程图像识别方法,其特征在于,所述步骤S4具体步骤如下:

S41以检测到的每个角点为中心截取 41×41 像素的角点邻域图像,形成角点样本集合,根据CBSM模型对部分角点样本进行人工分类标注,构造训练样本集和测试样本集;

S42提取角点邻域图像的网格特征和外围特征共128维特征,构造具有径向基核函数(Radial Basis Function,RBF)的SVM分类器,使用训练样本集对其进行训练并通过K折交叉验证选择最优SVM分类器参数。

8. 根据权利要求1所述流程图像识别方法,其特征在于,所述步骤S5流程图像识别分别

如下:

S51按照起始角点选择的优先级,在流程图中查找第一个未标注所属图元编号且未处理的角点作为起始角点,并将其标注为已处理角点;

起始角点的选择包括:R-1,E1-1,D*-1,R-2,Er-1,D*-3以及D*-4;

起始角点的优先级为:

不同类型间优先级:R-1=E1-1=D*-1>R-2=Er-1=D*-3=D*-4;

优先级相同时:按照从上到下,从左到右的位置顺序选择起始角点;

S52从起始角点出发,按照CBSM中的角点组合规则,寻找可能构成图元的角点组合,若未找到满足条件的角点组合则返回S51;

S53验证所获角点组合是否满足CBSM中给定的角点组合约束,若满足则在识别结果中添加新图元并对其进行编号,标注与相关角点的所属图元编号,若不满足返回S52;

S54采用OCR(Optical Character Recognition)技术识别图元对应位置处的文本信息,生成流程图的文本描述;

S55检测未标注的角点集合是否可能构成新图元,若可能构成新图元则返回S51。

一种流程图像识别方法

技术领域

[0001] 本发明属于图像识别技术领域,特别涉及一种流程图像识别方法。

背景技术

[0002] 现有计算机信息检索系统主要采用基于文本匹配的检索技术,但化工、制药、机械、信息技术等领域中的文档往往还包含大量流程图像,这些图像 蕴含丰富的信息,对于检索和查新至关重要。

[0003] 流程图可以直观地描述一个工作过程的具体步骤,具有重要的语义。由于现存文档中的流程图存在文图粘连和断边等情况导致流程图像识别效率不 高,因此需要一种能够新的流程图像识别方法,解决以上问题。

发明内容

[0004] 本发明的目的在于提供一种流程图像识别方法,通过基于角点特征的结构模型来识别流程图像。本方法包括建立一个基于角点的流程图结构模型 (CBSM),然后对流程图结构进行角点检测与角点分类,依据角点组合规则和组合约束来识别流程图结构元素。

[0005] 为了实现上述目的,本发明采用的技术方案是:

[0006] 一种流程图像识别方法能够基于角点特征描述和识别流程图结构,步骤包 括:

[0007] S1建立CBSM,定义角点类型、描述流程图结构元素的角点组合规则以 及角点组合约束包括:

[0008] S11将描述流程图结构元素的角点分为独立型和连接型两大类,命名各 类角点并建立流程图像角点分类表;

[0009] S12建立角点组合规则;

[0010] S13建立角点组合约束。

[0011] S2提取流程图像结构;

[0012] 对原始流程图像进行二值化、降噪处理,通过图文分割提取原始流程图像 中的结构图层,对提取的流程图结构进行单像素化。

[0013] S3对流程图结构进行角点检测和筛选,步骤为:

[0014] S31通过CSS算法利用曲率检测出与直线型元素相关的角点,过滤掉曲线 型结构元素上的圆角点和虚假角点。

[0015] 不同尺度 σ 下的曲率计算公式为:

[0016]
$$K(t, \sigma) = \frac{\dot{X}(t, \sigma)\ddot{Y}(t, \sigma) - \ddot{X}(t, \sigma)\dot{Y}(t, \sigma)}{[\dot{X}(t, \sigma)^2 + \dot{Y}(t, \sigma)^2]^{\frac{3}{2}}}$$
, 其中 t 是弧长参数, σ 是尺度参数,

$\dot{X}(t, \sigma)$ 、 $\dot{Y}(t, \sigma)$ 、 $\ddot{X}(t, \sigma)$ 、 $\ddot{Y}(t, \sigma)$ 分别为曲线 l 上一点的坐标 $x(t)$ 和 $y(t)$ 与一维 高斯核函数 $g(\mu, \sigma)$ 的一阶导数和二阶导数的卷积;

[0017] 圆角点判定方法为:

[0018] $R_c = \begin{cases} 1, & K(u) \leq T(u) \\ 0, & \text{other} \end{cases}$, 其中 u 为候选角点的位置参数, $K(u)$ 是候选角点的曲率,

$T(u)$ 为与角点支持域自适应的动态局部阈值,与候选角点 u 处的局部平均曲率成正比,当 $R_c=1$ 时表示角点为圆角点,给予滤除。

[0019] C_c is $\begin{cases} \text{true corner}, & \angle C_c \leq \theta_{\text{obtuse}} \\ \text{false corner}, & \angle C_c > \theta_{\text{obtuse}} \end{cases}$, 其中 C_c 为需要判定的候选角点, $\angle C_c$

为角点 C_c 的角, θ_{obtuse} 为真角点的最大钝角值, θ_{obtuse} 的经验取值为 162° , 当 $C_c > \theta_{\text{obtuse}}$ 时, C_c 为虚假角点。

[0020] S32通过Harris算法检测与曲线型元素相关的角点,首先求输入图像 I 对 应的 x 和

y 轴方向的一阶偏导 I_x 和 I_y ,利用高斯核 $G_{x,y}$ 对其平滑,并构造自相关矩阵 M , $M = \begin{bmatrix} \overline{I_x^2 I_x I_y} \\ \overline{I_x I_y I_y^2} \end{bmatrix}$,

其中, $\overline{I_x^2} = I_x^2 \otimes G$, $\overline{I_y^2} = I_y^2 \otimes G$, $\overline{I_x I_y} = I_x I_y \otimes G$, \otimes 是卷积算子, σ 是高斯平滑尺度。通过自相关矩阵的特征值 α, β ($\alpha \geq \beta$) 来构造角点响应 R 并判决角点, $R = \alpha\beta - k(\alpha + \beta)^2$, 其中, k ($k = 0.04$) 是错误角点响应抑制常数。

[0021] S33采用距离筛选法对检测出的冗余角点进行筛选。计算指定范围内所有点之间的欧式距离,在欧几里得空间中,如点 $x = (x_1, \dots, x_n)$ 和 $y = (y_1, \dots, y_n)$ 之间的距离为:

$$[0022] \quad d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

[0023] 当角点间的距离小于指定阈值时,可删掉其中之一以减少冗余。

[0024] S4得到步骤S3中角点的邻域特征并进行角点分类器的训练,对流程图中角点进行

分类:

[0025] S41以检测到的每个角点为中心截取 41×41 像素的角点邻域图像,提取角点邻域图像的网格特征和外围特征并将角点样本分类标注;

[0026] S42使用支持向量机SVM分类器训练样本集,采用径向基核函数(Radial Basis Function, RBF)作为SVM分类核函数;利用MATLAB中的LIBSVM工具箱来实现SVM多分类,引入惩罚因子 c 对其进行惩罚,通过参数调优设置惩罚因子 c ,使得数据在高维特征空间中的线性可分度最大。

[0027] S43采用K-折交叉验证(K-fold Cross Validation, K-CV)将原始数据分成均分为 K 组,将每个子集数据分别做一次验证集,其余的 $K-1$ 组子集数据作为训练集,用这 K 个模型最终验证集的分类准确率的平均数作为此K-CV下分类器的性能指标。

[0028] S5采用基于CBSM结构元素识别算法对流程图像图元识别,步骤为:

[0029] S51按照起始角点选择的优先级,在流程图中查找第一个未标注所属图元编号且未处理的角点作为起始角点,并将其标注为已处理角点;

[0030] 起始角点的选择包括: $R-1, E1-1, D*-1, R-2, Er-1, D*-3$ 以及 $D*-4$ 。

[0031] 起始角点的优先级为:

[0032] 不同类型间优先级: $R-1 = E1-1 = D*-1 > R-2 = Er-1 = D*-3 = D*-4$;

- [0033] 优先级相同时：按照从上到下，从左到右的位置顺序选择起始角点。
- [0034] S52从起始角点出发，按照CBSM中的角点组合规则，寻找可能构成图元的角点组合，若未找到满足条件的角点组合则返回S51；
- [0035] S53验证所获角点组合是否满足CBSM中给定的角点组合约束，若满足则在识别结果中添加新图元并对其进行编号，标注与相关角点的所属图元编号，若不满足返回S52；
- [0036] S54采用OCR (Optical Character Recognition) 技术识别图元对应位置处的文本信息，生成流程图的文本描述；
- [0037] S55检测未标注的角点集合是否可能构成新图元，若可能构成新图元则返回S51。
- [0038] 与现有技术相比，本发明采用基于角点特征的结构模型的流程图像识别方法，对流程图结构元素提供了基于角点的新的定义方式，可以有效解决文图粘连和断边对专利流程图识别的影响，对整体流程图的识别快速有效。

附图说明

- [0039] 图1是本发明的流程示意图。
- [0040] 图2是本发明流程S3示意图。
- [0041] 图3是本发明流程S5示意图。
- [0042] 图4是本发明实验获取到的36种角点示意图。
- [0043] 图5是本发明流程S6示意图。

具体实施方式

- [0044] 为使本发明的目的、技术方案和优点更加清楚，下面结合实施例、附图对本发明作进一步描述。显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。
- [0045] 如图1所示，本发明的识别方法的步骤包括：
- [0046] S1建立一个基于角点的流程图结构模型 (CBSM)，定义描述流程图结构元素的角点组合规则以及角点组合约束；
- [0047] S2提取流程图结构；
- [0048] S3对流程图结构进行角点检测和筛选；
- [0049] S4得到步骤S3中角点的邻域特征并进行角点分类器的训练，对角点进行分类；
- [0050] S5基于CBSM对流程图结构元素进行识别；
- [0051] S6文本识别以及结构描述生成。
- [0052] 其中步骤S1包括：
- [0053] S11将流程图结构元素分为两大类，命名各个独立型角点和连接型角点，建立如表1所示流程图像角点分类表，表1中上半部分为独立型角点，表1下半部分为连接型角点；
- [0054] 表1流程图像角点分类

[0055]

结构元素名称	结构元素	角点及命名			
矩形					
折线					
菱形					
椭圆					
矩形-连接线					
连接线-连接线					
菱形-连接线					
椭圆-连接线					

[0056] S12根据图元与角点组合之间的关系,建立组合规则。

[0057] 将基本图元的角点组合规则定义为三类关系:角点组合关系,角点间相对位置关系,角点的几何关系。通过角点组合判断得到结构图元的识别结果。

[0058] 角点组合关系定义为确定类型角点的并集。

[0059] 角点间相对位置关系的定义为:假设 $C_i, C_j (i \neq j)$ 表示两个不同类型的角点, $x(C_i)$ 表示角点 C_i 横坐标, $y(C_i)$ 表示角点 C_i 纵坐标。定义角点 C_j 相对于角点 C_i 的相对位置可以有8种表示,分别为左上(u1),左(l),左下(d1),下(d),右下(dr),右(r),右上(ur),上(u),每种位置关系对应一个判定表达式。

[0060] 例如, $r(C_i, C_j)$ 表示 C_j 在 C_i 的右方位置,满足 $x(C_i) < x(C_j)$; $d(C_i, C_j)$ 表示 C_j 在 C_i 下方位置,满足 $y(C_i) < y(C_j)$; $d1(C_i, C_j)$ 表示 C_j 在 C_i 的左下方位置,满足 $(x(C_i) > x(C_j)) \wedge (y(C_i) < y(C_j))$ 。其中, $r(C_i, C_j)$ 等价于 $l(C_j, C_i)$, $d(C_i, C_j)$ 等价于 $u(C_j, C_i)$, $dr(C_i, C_j)$ 等价于 $u1(C_j, C_i)$, $d1(C_i, C_j)$ 等价于 $ur(C_j, C_i)$, $d(C_i, C_j)$ 等价于 $u(C_j, C_i)$, $r(C_i, C_j)$ 等价于 $l(C_j, C_i)$ 。

[0061] 在流程图结构元素中,角点间几何关系定义为水平关系(H)或者垂直(V)关系。 $H(C_i, C_j)$ 表示 C_i 与 C_j 处于同一条水平线上,满足 $y(C_i) = y(C_j)$; $V(C_i, C_j)$ 表示 C_i 与 C_j 处于同一条垂直线上,满足 $x(C_i) = x(C_j)$ 。其中, $H(C_i, C_j)$ 等价于 $H(C_j, C_i)$, $V(C_i, C_j)$ 等价于 $V(C_j, C_i)$ 。

[0062] R图元的角点组合表示为 $R-1 \wedge R-2 \wedge R-3 \wedge R-4$ 。

[0063] R图元的角点相对位置表示如下:

[0064] $r(R-1, R-2) \wedge d(R-1, R-3) \wedge dr(R-1, R-4) \wedge d1(R-2, R-3) \wedge d(R-2, R-4) \wedge r(R-3, R-4)$

[0065] R图元的角点间几何关系表示如下:

[0066] $H(R-1, R-2) \wedge V(R-1, R-3) \wedge H(R-3, R-4) \wedge V(R-2, R-4)$

[0067] E图元的角点组合表示为 $E1-1 \wedge E1-2 \wedge Er-1 \wedge Er-2$ 。

[0068] E图元的角点之间相对位置表示如下：

[0069] $r(E1-1, Er-1) \wedge d(E1-1, E1-2) \wedge dr(E1-1, Er-2) \wedge dl(Er-1, E1-2) \wedge d(Er-1, Er-2) \wedge r(E1-2, Er-2)$

[0070] E图元的角点间几何关系表示如下：

[0071] $H(E1-1, Er-1) \wedge V(E1-1, E1-2) \wedge H(E1-2, Er-2) \wedge V(Er-1, Er-2)$

[0072] D图元的角点组合表示为 $D*-1 \wedge D*-2 \wedge D*-3 \wedge D*-4$, D^* -表示D-型 (独立型) 角点或Db-型 (连接型) 角点。

[0073] D图元的角点相对位置表示如下：

[0074] $d(D*-1, D*-2) \wedge dl(D*-1, D*-3) \wedge dr(D*-1, D*-4) \wedge ul(D*-2, D*-3) \wedge ur(D*-2, D*-4) \wedge r(D*-3, D*-4)$

[0075] D图元的角点间几何关系表示如下：

[0076] $H(D*-3, D*-4) \wedge V(D*-1, D*-2)$

[0077] 定义描述流程图结构元素的角点组合所满足的约束, 具体包括角点类型 和数量约束、角点间的几何约束以及角点组合构成的图元的宽高约束。分别 定义如下。

[0078] 各图元的角点组合中角点类型和数量的约束如下：

[0079]	R图元	$\begin{cases} \text{独立型角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 2; \\ \text{同一类型角点数量} < 2; \end{cases}$	E图元	$\begin{cases} \text{独立型角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 1; \\ \text{同一类型角点数量} < 2; \end{cases}$
[0080]	D图元	$\begin{cases} \text{角点类型数量} \geq 3; \\ \text{连接型角点数量} \geq 2; \\ \text{同一类型角点数量} < 2; \end{cases}$		

[0081] 角点组合中的角点间几何约束指的是两个角点在一定像素范围内呈水平 关系 (H) 或者垂直关系 (V)。分别定义为： $H(C_i, C_j) :: |y(C_i) - y(C_j)| < d$, $V(C_i, C_j) :: |x(C_i) - x(C_j)| < d$, d 为修正在角点检测过程中出现的位置偏差所 设置的松弛因子, 经验取值为5px。

[0082] 角点组合所构成图元的宽和高约束如下所述, 若流程图像的整体宽和高 分别为 W, H , R_{width} 表示R图元的宽, R_{height} 表示R图元的高, 则：

[0083] R图元: $R_{width} < W \wedge R_{height} < \frac{H}{2}$;

[0084] E_{width} 表示E图元的宽, E_{height} 表示E图元的高, 则：

[0085] E图元: $E_{width} < \frac{W}{2} \wedge E_{height} < \frac{H}{3}$;

[0086] D_{width} 表示D图元的宽, D_{height} 表示D图元的高, 则：

[0087] D图元: $E_{width} < W \wedge E_{height} < \frac{H}{2}$;

[0088] 其中步骤S5包括：

[0089] S51按照起始角点选择的优先级, 在流程图中查找第一个未标注所属图 元编号且未处理的角点作为起始角点, 并将其标注为已处理角点；

[0090] 起始角点的选择包括: R-1, E1-1, D*-1, R-2, Er-1, D*-3以及D*-4。

- [0091] 起始角点的优先级为：
- [0092] 不同类型间优先级： $R-1=E1-1=D*-1>R-2=Er-1=D*-3=D*-4$ ；
- [0093] 优先级相同时：按照从上到下，从左到右的位置顺序选择起始角点。
- [0094] S52从起始角点出发，按照CBSM中的角点组合规则，寻找可能构成图元的角点组合，若未找到满足条件的角点组合则返回S51；
- [0095] S53验证所获角点组合是否满足CBSM中给定的角点组合约束，若满足则在识别结果中添加新图元并对其进行编号，标注与相关角点的所属图元编号，若不满足返回S52；
- [0096] S54采用OCR(Optical Character Recognition)技术识别图元对应位置处的文本信息，生成流程图的文本描述；
- [0097] S55检测未标注的角点集合是否可能构成新图元，若可能构成新图元则返回S51。
- [0098] 图2中空角点为对流程图像进行检测得到的直线型角点；实心角点为对流程图像进行检测并经过距离筛选得到的曲线型角点；其中右下方矩形由于文字超出形成断边情况，断边处的角点同样被识别。
- [0099] 图3是采用本发明使用从CLEP-IP 2012公开数据集中获取的50张流程图和从网络爬取的50张流程图共100张作为实验对象，对专利流程图中角点进行标注后获取到的角点种类，其中上部分对应表1中所定义的流程图像中20种角点，下部分为是实际情况中出现其他角点种类。
- [0100] 采用本发明，从实验流程图中截取2600张角点作为角点样本，采用LIBSVM进行角点的多分类，对于SVM参数的优化，本实验选择网络搜索优化方式，利用LIBSVM中grid.py参数工具，可以自动进行参数寻优，最终得到最佳参数。
- [0101] 利用参数工具得到的c,g值在SVM中建模，对角点类型进行预测。将实验数据集均分为5组，每组角点520张。各类角点识别率统计如表3所示，实验中可分为8类角点，N代表每一类角点的个数，P为角点的查准率，R为查全率。根据每一组训练集来进行测试集角点类型的检测，并计算查全率和查准率以及每一角点类型的平均值。从实验结果总结分析，对测试集角点分类结果较理想，对5组中8类角点的实验结果求均值得到查准率为89.1%，查全率为91.6%。
- [0102] 表3各组角点分类结果
- [0103]

角点类型	1		2		3		4		5		平均		第1组
	P/%	R/%	P/%	R/%	P/%	R/%	P/%	R/%	P/%	R/%	P/%	R/%	Top3R/%
R	95.8	91.9	96.2	93.3	98.4	91.3	98.2	92.1	97.9	91.3	97.3	91.9	100
Rb	90.6	95.3	87.6	96.0	86.6	92.6	87.8	89.5	95.8	90.5	89.7	92.4	94.3
L	92.5	96.5	93.5	90.3	95.5	92.9	92.3	95.8	95.4	93.1	93.8	93.6	97.8
E	88.3	90.7	85.3	91.7	89.7	93.5	81.2	91.6	81.2	91.6	85.1	91.8	89.9
Eb	81.3	93.7	90.7	94.2	89.9	88.6	86.9	80.4	87.6	94.9	87.3	91.1	87.5
D	91.2	83.6	86.2	92.5	81.2	87.7	76.2	85.1	94.2	92.8	85.8	89.1	96.7
Db	98.3	89.2	91.3	89.1	79.3	93.2	80.1	93.8	92.7	95.8	88.3	92.8	98.3
A	84.7	91.4	88.9	92.4	89.6	91.5	91.7	90.0	69.6	93.2	84.9	91.9	95.3
平均	90.3	91.5	90.0	92.4	88.8	91.4	86.8	89.9	89.3	92.9	89.1	91.6	-

[0104] 图4为流程图被识别的基本图元，如图的实施例中，流程图基本图元矩形、菱形、圆角矩形均被识别；对于右下角出现断边情况的矩形也被准确识别。将实验所得流程图的文本结果和CLEP-IP 2012公开数据集中提供的ground truth进行对比验证。在分类过

程中,根据图元本身类型(真/假)以及识别结果(真/假),可能出现以下几种情况:本身是真,识别结果也为真(TP);本身是真,识别结果为假(FN);本身是假,识别结果为真(FP)。统计E、D、R图元的数量、误识别率、查全率(R)以及查准率(P),如表4所示。

[0105] 表4图元识别率统计

[0106]	图元类型	<i>N</i>	<i>TP</i>	<i>FP</i>	<i>FN</i>	<i>R</i> (%)	<i>P</i> (%)
	E	158	121	35	2	98.4	77.6
	R	573	557	11	5	99.1	98.1
	D	123	114	19	0	100	92.7
	总计	854	792	65	7	99.1	92.4

[0107] 图5是实施例流程图最终识别结果的形式描述。

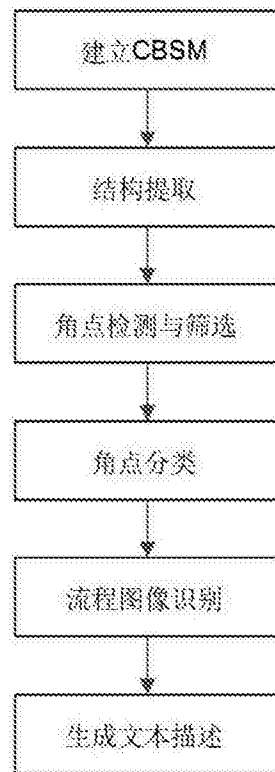


图1

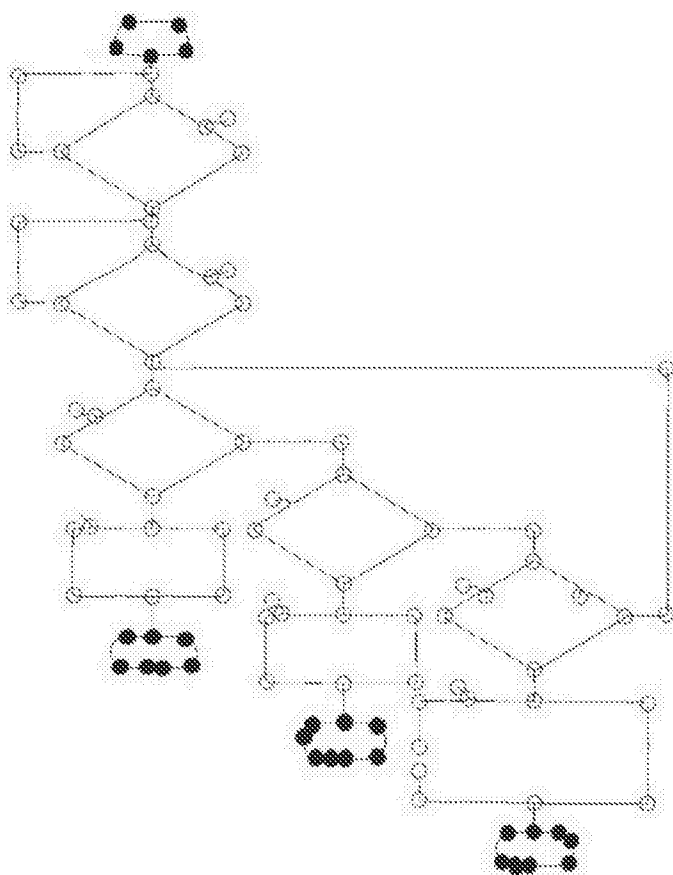


图2

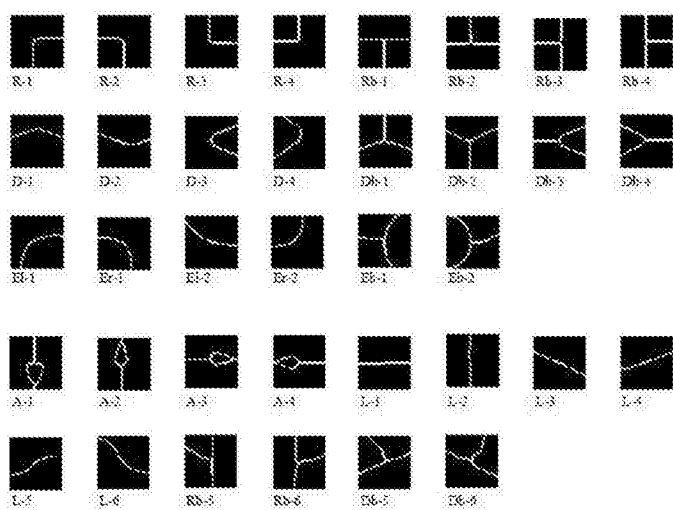


图3

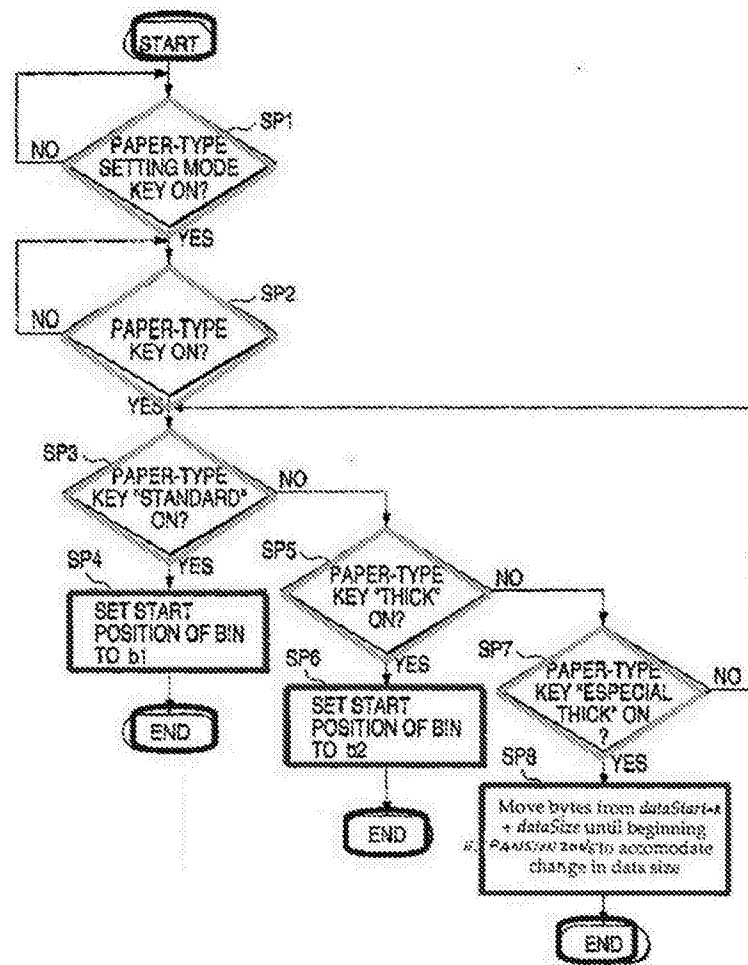


图4

NO01 diamond PAPER-TYPESETTING MODE KEY ON?
 NO02 diamond / -PAPER-TYPE KEY ON?
 NO03 diamond / PAPER-TYPE KEY "STANDARD" ON?
 NO04 diamond / PAPER-TYPE KEY "THICK" ON?
 NO05 diamond SET START POSITION OF BENTO 131
 NO06 rectangle SET START POSITION OF BENTO
 NO07 rectangle Move b7/- 6/om dom- Stam* / + dataS - z, until 60gning
 NO08 oval STAKT
 NO09 oval END
 NO10 oval END
 NO11 oval END

图5