

基于多候选的数学公式识别系统

郭育生^{1,2} 黄磊¹ 刘昌平¹

¹(中国科学院自动化研究所 北京 100080)

²(中国科学院研究生院 北京 100080)

(yusheng.guo@ia.ac.cn)

A Multi-Candidate Mathematical Expression Recognition System

Guo Yusheng^{1,2}, Huang Lei¹, and Liu Changping¹

¹(*Institute of Automation, Chinese Academy of Sciences, Beijing 100080*)

²(*Graduate University of Chinese Academy of Sciences, Beijing 100080*)

Abstract A multi-candidate mathematical expression (ME) recognition system is proposed. The system includes three main components: image preprocessing, symbol segmentation and structure analysis. During the symbol segmentation period, a three-stage segmentation method based on dynamic programming (DP) is proposed. In the initial segmentation based on DP algorithm, the ME image is segmented into several blocks. In the vertical segmentation, each block is segmented into some blobs. In the horizontal segmentation based on DP algorithm, every blob is segmented into symbols. During the structure analysis period, hierarchical structure is adopted to analyze structure of ME. The hierarchical structure analysis method consists of three steps, i.e., matrix analysis, sub-expression analysis and script expression analysis. In matrix analysis (sub-expression analysis), an ME is decomposed into several basic matrixes (basic sub-expressions) and some sub-expressions (script expressions) by reconstructing the ME (sub-expression) global structure, and then every basic matrix (sub-expression) is analyzed from bottom to up. In script analysis, a graph rewriting algorithm is adopted to build script relation trees among symbols within a script expression. A spatial relation model is built to calculate spatial relations' confidence between two symbols. The experiments are implemented on a database with 3268 ME images and the results show that the proposed system works well. Top-1 ME recognition accuracy reaches 78.2%.

Key words multi-candidate; printed mathematical expression; dynamic programming; hierarchical structure; spatial relation model

摘要 提出了一种基于多候选方法的数学公式识别系统。该系统主要包括公式图像预处理、多候选公式符号分割和多候选公式结构分析 3 个部分。在公式符号切分中,使用 3 次动态规划方法对公式图像进行多候选公式符号切分。在公式结构分析中,采用层次结构方法多候选分析公式符号间的结构关系,然后使用 LaTeX 格式和 MathType 格式表示数学公式的识别结果。为了确定符号间的空间位置关系,建立了符号的空间关系模型。在 3268 个公式图像组成的测试集上取得了 78.2% 的公式分析正确率。

关键词 多候选;印刷体数学公式;动态规划;层次结构;空间关系模型

中图法分类号 TP391

科技工程文献中往往存在着大量数学公式,为了重用这些科技工程文献,OCR系统不仅需要具有文字识别的能力,还需要具有数学公式识别的能力。研究人员围绕公式识别进行了大量的研究,但迄今为止还没有实用的公式识别产品。

在过去的几十年中,研究人员提出了多个数学公式识别系统^[1-2]。Lee等人提出的系统能够处理简单的数学公式并取得了一定的效果,但该系统不能处理多行数学公式以及比较复杂的单行公式且该系统的测试图像集较小^[3]。Fateman等人设计的系统原型能够将无噪音的数学公式图像翻译成Lisp表达式,但是实际系统只能识别固定格式的积分表^[4]。Okamoto等人实现的系统针对实际扫描的公式图像取得了较好的分析效果并支持识别性能的自动评测,但该系统不能分析带有两个以上符号的修饰符表达式,并且对于上下标和矩阵支持得不好^[5-8]。靳简明实现的系统在结构分析中取得较好的效果,但没有提到分割性能,仍未达到实际应用的程度^[9]。

本文提出了一种基于多候选方法的数学公式识别系统,图1为系统的流程图。该系统主要分为4个部分:预处理;多候选数学公式符号分割;多候选公式符号结构分析;公式识别结果转换。在多候选数学公式符号分割中,为了分割公式中的符号包括粘连符号和断裂符号,采用3次动态规划方法对公式图像进行多候选分割。在多候选数学公式结构分析中,首先采用层次结构分析方法分析数学公式中的矩阵和子表达式,然后使用图搜索的方法多候选分析角标表达式并使用文法删除不合法的候选。为了计算符号间的空间关系,本系统建立了空间关系模型。

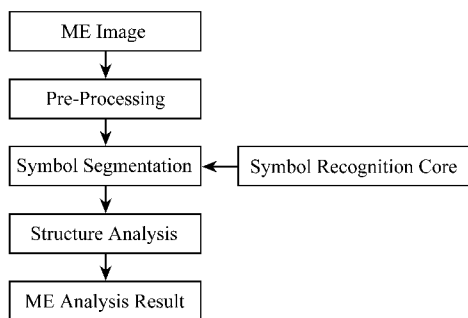


Fig. 1 ME recognition framework.

图1 数学公式识别系统框架

1 预处理

在数学公式识别系统中,预处理主要包括噪声滤除、图像倾斜矫正和图像二值化等。图2为预处理

后的公式图像,本文以该图像为例说明数学公式识别系统的识别过程。

$$\left(\begin{array}{c|c} p_j^2+t & 1 \\ \hline 1 & n \\ n+1+t & \\ \hline 0 & 1 \end{array} \right)$$

Fig. 2 An example of ME image.

图2 公式图像举例

2 多候选数学公式符号分割

数学公式符号分割就是从数学公式图像中分割出带有属性的公式符号串。公式符号分割主要包括初始分割公式图像,竖直方向分割公式子图,水平方向分割公式子图和函数名合并4个部分。

2.1 初始分割公式图像

在公式图像初始分割中首先使用动态规划方法搜索图像中可能的分割路径,然后根据可能的分割路径分割公式图像获得公式子图,最后公式图像的初始分割结果如图3所示:

Fig. 3 Initial segmentation result.

图3 初始分割阶段公式图像分割结果

2.2 竖直方向上多候选分割公式子图

在公式图像中,公式图像中的符号排列不仅可以按照水平邻接关系排列,还可以按照垂直关系排列,有时甚至是按照包含关系排列。这样对于公式图像很难仅通过水平方向分割就分割出正确的分割结果。因此在初始分割公式图像后,再从竖直方向上使用动态规划方法多候选分割公式子图。

在竖直方向分割子图中,首先使用动态规划方法搜索可能的分割路径,然后在带有拒识模型的识别核心提供的符号识别可信度基础上使用动态规划方法合并路径间的图像并形成竖直方向的分割结果,

Fig. 4 Top-1 vertical segmentation result.

图4 竖直方向公式分割结果(1选)

最后垂直方向分割公式图像的第 1 候选结果如图 4 所示。

2.3 空间关系模型

在数学公式中,公式符号之间存在一定的结构关系,为了计算这些关系的类别以及相应的可信度,本节建立符号间的空间关系模型。

首先根据公式中符号之间的结构关系将空间关系分为 C 类,并抽取 D 维的拓征,然后使用 GMM 模型建立空间关系分类的模型,再使用 EM 算法和 MCE 算法训练 GMM 模型的参数,最后使用训练好的 GMM 模型给识别空间关系分类。

2.3.1 特征抽取

空间关系模型定义空间关系共有 C ($C = 7$) 类即定义 7 种类型的空间关系,分别为:水平、上方、下方、右下角、右上角、重叠和包含。

定义特征的维数为 D ($D = 14$) 即本文中使用的特征维数为 14,这 14 个特征具体为:符号宽度 (w_1, w_2),符号高度 (h_1, h_2),边框的中心距离 ($\Delta x, \Delta y$),边框的中心角度 ϕ ;符号的中心距离 ($\Delta x', \Delta y'$),符号中心角度 ϕ' ;符号的 SHSI 分 ($S_x = \frac{\Delta x}{W}, S_y = \frac{\Delta y}{H}$);符号的边框面积 $m_1/M, m_2/M$ 。其具体含义如图 5 所示。其中:

$$m_1 = w_1 \times h_1, m_2 = w_2 \times h_2, M = W \times H.$$

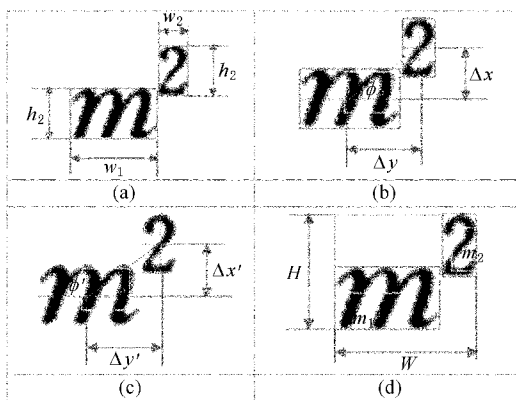


Fig. 5 Characters of spatial relation model. (a) w_1, w_2, h_1, h_2 ; (b) $\Delta x, \Delta y, \phi$; (c) $\Delta x', \Delta y', \phi'$; and (d) m_1, m_2, M, W, H .

图 5 空间关系特征。(a) w_1, w_2, h_1, h_2 ; (b) $\Delta x, \Delta y, \phi$; (c) $\Delta x', \Delta y', \phi'$; (d) m_1, m_2, M, W, H

2.3.2 基于空间关系的 GMM 模型

首先使用 Gaussian Mixture Model(GMM)建立空间关系的模型,再使用 EM^[10]算法和 MCE^[11]算法训练 GMM 模型,最后用训练后的 GMM 模型识别空间关系。

2.4 水平方向多候选分割公式图像

在垂直方向分割中得到的子图可能是 1 个符号形成的图像,也可能是多个符号形成的图像,因此本节再次使用动态规划从水平方向上多候选分割子图。首先使用动态规划搜索可能的图像分割路径,然后在带有拒识模型的识别核心提供的识别可信度和结合 bi-gram 信息的空间关系模型提供的空间关系可信度基础上,使用动态规划算法多候选合并子图最后形成公式符号串。最后水平方向分割公式子图的结果如图 6 所示:

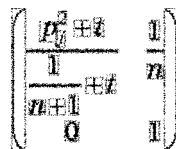


Fig. 6 Top-1 horizontal segmentation result.

图 6 水平方向公式分割结果(1 选)

在数学公式中存在许多常见的函数名如 $\sin, \lg, \text{cov}, \text{inf}, \text{sup}, \text{lim}, \text{exp}, \text{min}, \text{max}, \text{mod}, \text{rad}, \text{arg}, \text{det}$ 等,这些函数名在数学公式中起着特殊的作用,因此在这里使用 Bayesian 后验概率的方法进行合并函数名。最后公式符号的第 1 候选分割结果如图 6 所示。其中每个分割得到的符号都具有多个识别 ID 和相应的识别可信度,如图 6 中最左边的 '(' 识别结果见表 1 所示:

Table 1 Symbol Recognition Result

表 1 符号识别结果

ID	Confidence
(0.46
[0.39

3 数学公式符号多候选结构分析

基于层次结构的公式结构分析方法共分为 3 个阶段:多候选矩阵分析阶段、多候选子表达式分析阶段和多候选角标分析阶段。在矩阵分析和子表达式分析阶段,使用空间关系模型建立公式的层次结构,然后分析公式的每一层。在多候选角标分析阶段,使用带有拒识的空间关系模型建立符号间的角标关系图,然后使用图搜索的方法多候选搜索角标关系树,并使用文法剪枝获得 Top-N 个分析候选。

3.1 多候选矩阵分析

在多候选角标分析中,首先按照符号识别候选中是否包含定界符展开符号候选,然后对左定界符

和右定界符进行多候选配对形成定界符组,再通过空间关系模型建立定界符组之间的层次结构关系,最后对每一个定界符组依次进行符号搜索,行分析和列分析,最后得到矩阵分析结果。

3.1.1 多候选矩阵定界符配对及定界符组关系树

在每一个候选中搜索左定界符和右定界符。在搜索得到左定界符和右定界符后,通过空间关系模型产生定界符关系图,再搜索 Top-N 组配对结果。最后定界符配对的第 1 候选结果如图 7 所示。在图中浅色矩形框为配对后的定界符组边框。

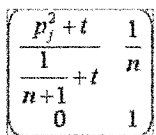


Fig. 7 Top-1 delimiters grouping result.
图 7 定界符组第 1 候选配对结果(1 选)

在定界符配对后使用空间关系模型建立定界符组之间的关系图,然后在关系图中搜索定界符组之间的关系树即矩阵分析阶段的定界符组层次结构关系,最后根据定界符组的层次结构关系依次分析每个定界符组。

3.1.2 多候选搜索矩阵内的符号

在矩阵内符号搜索中,首先使用空间关系模型计算每一个符号与定界符组的空间位置关系及其可信度,然后根据计算得到的空间关系和相应的可信度确定该符号位于定界符组内的可信度,最后获得定界符组的符号。在图 7 中除了定界符外所有符号都是定界符组内的符号。

3.1.3 多候选矩阵行分析

多候选矩阵行分析首先在垂直方向进行轮廓投影并初始分割矩阵行,然后多候选合并矩阵行并取 Top-N 个分析候选。最后行分析的结果如图 8 所示,其中浅色矩形框为行边框。

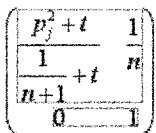


Fig. 8 Top-1 matrix rows analysis result.
图 8 矩阵行分析结果(1 选)

3.1.4 多候选矩阵列分析

在矩阵列分析中首先计算矩阵元素的高度并根据矩阵元素高度估计矩阵的列间距离,然后使用水平方向轮廓投影方法多候选分割矩阵列。最后矩阵

列分析的第 1 候选结果如图 9 所示,其中浅色矩形框为矩阵列的边框。

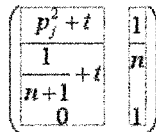


Fig. 9 Top-1 matrix column analysis result.
图 9 矩阵列分析结果(1 选)

最后矩阵分析阶段的分析结果如图 10 所示,其中 Sub₁, Sub₂ 均为子表达式,对应节点上的子表达式为一带有属性的符号串。

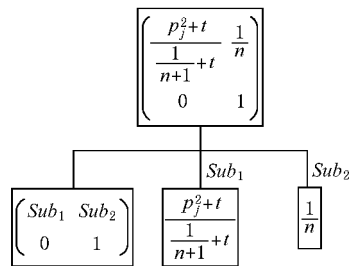


Fig. 10 Top-1 matrix analysis results.
图 10 矩阵阶段公式分析结果(1 选)

3.2 多候选子表达式分析

在子表达式分析中,首先按照符号识别结果中是否包含控制符(表达式的核心符号,如图 11 中的分数线)展开候选,然后在每一个候选中建立控制符之间的层次结构关系,最后依次分析每一个简单子表达式。本节以图 10 中的 Sub₁ 所示的子表达式图像为例说明多候选子表达式层的分析方法。

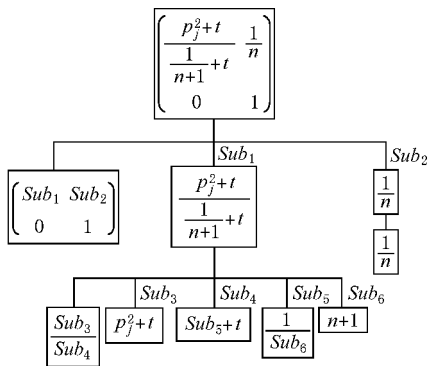


Fig. 11 Top-1 sub-expression analysis result.
图 11 子表达式阶段公式分析结果(1 选)

3.2.1 子表达式层次结构关系

在建立子表达式控制符的层次结构关系中,首先利用空间关系模型建立控制符之间的关系图,然后多候选搜索控制符的层次结构关系。在 Sub₁ 中,

控制符的层次结构关系为 : S_1 Below $\{S_2\}$, 其中 S_1 表示第 1 个符号(符号顺序按照左边框从左至右排列. 若左边框相同, 则按照从上到下排列相同左边框的符号, 在图 11 中 Sub_1 中第 1 个符号为一分数线).

3.2.2 子表达式分析

首先建立符号与控制符之间的空间关系图, 然后多候选搜索控制符的元素, 再按照可信度取 Top-N 个子表达式分析候选. 最后子表达式阶段的分析结果如图 11 所示, 其中 Sub_3, Sub_4, Sub_6 有待进一步的分析.

3.3 多候选角标分析

多候选角标分析是指分析前面分析得到的每一个待分析的角标表达式以便得到公式的最终分析结果. 这里以图 11 中的角标表达式 Sub_3 为例说明角标分析的方法.

在角标层的空间关系模型中, 符号之间的关系只有 4 种: 上标、下标、水平和拒识. 其中拒识关系对应于 7 种空间关系中的其他 4 种. 基于空间关系模型建立两两符号间的关系图, 最后获得两两符号的空间关系图.

在获得符号间的角标关系图后, 使用图搜索的方法搜索 Top-N 个角标关系树, 如 $p \supset \{2\}_{sub \{2\}}$. 其中 \sup 表示上标关系, sub 表示下标关系, hor 表示水平关系.

空间关系可信度不仅与空间关系类型和相应的可信度有关, 还与相邻符号的文法可信度有关, 因此在大量数据统计的基础上建立符号间的文法信息并使用文法信息计算相邻符号间的角标关系可信度并使用平滑算法^[12]进行平滑. 根据角标表达式分析可信度选取 Top-N 个候选. 对所有的角标表达式分析后, 数学公式的分析结果如图 12 所示:

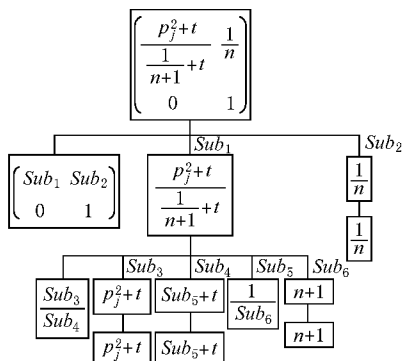


Fig. 12 Top-1 ME analysis result.

图 12 公式分析结果(1 选)

在获得数学公式结构分析结果(结构关系树如图 12 所示)后, 用 LaTeX 格式和 MathType 格式分别表示这些结构关系树, 如图 13 和图 14 所示:

```

\left( \begin{array}{r}
\frac{p_j^2+t}{n+1+t} \quad \frac{1}{n} \\
\frac{1}{n+1} \quad 1
\end{array} \right)

```

Fig. 13 ME recognition result with LaTeX format.

图 13 公式识别结果表示格式: LaTeX 格式

$$\left(\begin{array}{r} \frac{p_j^2+t}{n+1+t} \quad \frac{1}{n} \\ \frac{1}{n+1} \quad 1 \end{array} \right)$$

Fig. 14 ME recognition result with MathType format.

图 14 公式识别结果表示格式: MathType 格式

4 实验结果及分析

4.1 实验结果

为了比较全面地反映公式识别系统的性能, 本文扫描的图像来源不仅有数学书籍还有期刊, 不仅有高中书籍、高中数学手册, 还有现代数学手册和研究生数学课本, 这些书籍一共有 12 本: 现代数学手册——经典数学卷, 现代数学手册——计算数学卷, 现代数学手册——随机数学卷, 现代数学手册——现代数学卷, 应用随即过程课本, 中科院研究生院期刊、数学期刊、高中数学课本、高中数学手册、两本高中数学试卷集和一本高中数学参考书.

从这些书籍中以 300dpi 的分辨率一共扫描了 3420 多页包含数学公式的灰度图像, 并手工扣取了 7400 多个比较复杂(扣取的公式中符号数不少于 20 个)的公式图像并标注了 3268 个公式作为测试公式集. 在测试集中包含了常见的 8 种基本类型的数学表达式及其复合嵌套形成的数学公式: 由定界符() [] { } || || 形成的矩阵; 不带根指数的根式, 带有根指数的根式; 分式, 组表达式(求和表达式、连乘表达式、与运算、并运算、帽子表达式); 函数名表达式(求极限函数名、求最小最大函数名等); 修饰符表达式(矢量表达式、上画线表达式、点修饰符表达式等); 堆叠表达式; 角标表达式(上标、下标、水平表达式).

在本文的数学公式识别系统中, 每幅公式图像的识别所需平均时间为 2.65s (PIV 3.0GHz). 在

3268 个公式中,公式符号分割正确率 1 选为 95.6%,5 选正确率为 97.2%。在标注分割结果上,数学公式结构分析正确率 1 选为 92.1%,5 选为 93.4%。公式完全分析正确率 1 选为 78.2%,5 选为 82.5%。结构分析正确率是在标注分割基础上获得的正确率,公式完全分析正确指公式中的符号识别完全正确同时符号之间的结构分析也完全正确,图 15 为部分识别完全正确的公式。

$$= \frac{1}{2^s} \left(\frac{1}{x_1 x_2 \dots x_s} \times \frac{2^k - 1}{2^k - 1} + \frac{b_1 b_2 \dots b_s}{\underbrace{[1 \dots 1]}_k} \right)$$

$$d = \frac{\sqrt[3]{A + \sqrt{B}}}{\begin{vmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ m_1 & n_1 & p_1 \\ m_2 & n_2 & p_2 \end{vmatrix}}$$

$$\frac{n_h}{n} = \frac{\frac{W_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{c_h}}} = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^L \frac{N_h S_h}{\sqrt{c_h}}}$$

Fig. 15 Scanned image of ME analyzed perfectly.

图 15 正确公式分析示例

4.2 实验结果分析

实验结果表明,本文的基于多候选的数学公式识别系统能够有效地分析各种类型的数学公式。在测试数据集上公式分析正确率取得了 1 选 78.2% 的正确率。

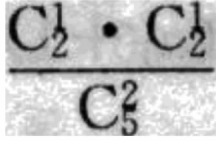
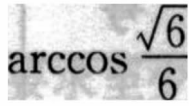
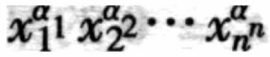
在 3 次动态规划方法分割公式符号中,竖直方向分割子图和水平方向分割子图方法取得了较好的分割效果,符号分割 1 选正确率为 95.6%,单候选的分割方法符号分割正确率 1 选为 93.2%,而基于连通域的分割方法符号分割正确率仅为 86.2%。

多候选公式结构关系分析中,每一阶段都保存多个可能分析分割候选,因而每一阶段的分析候选中包含正确分析结果的概率高于单候选的分析结果,因此多候选的公式结构分析正确率高于单候选的结构分析正确率。在使用本文的分割方法的测试集中,多候选的结构分析算法获得的公式分析 1 选正确率为 78.2%,单候选的结构分析算法(即每个阶段的分割结果均只保留第 1 候选)获得的公式分析正确率仅为 69.3%,基于基准线的公式结构分析方法获得的正确率为 59.6%。

虽然本文的公式识别系统取得了较好的公式识别效果,但仍然存在一些识别错误,如表 2 所示。引起公式分析错误的原因主要有:符号的粘连形成的

分析错误;符号断裂形成的分析错误;多重角标引起的分析错误等。

Table 2 Scanned Image of ME Analyzed Wrongly
表 2 错误分析的公式示例

ME Image	Analysis Result
	$\frac{L2 \cdot C_i}{C \%}$
	$\arcco. \cdot h \frac{\sqrt{6}}{6}$
	$x_1^\alpha x_2^\alpha \dots x_n^\alpha$

5 结论和进一步研究

为了识别科技工程文献中的数学公式,本文基于多候选的方法提出了一种数学公式识别系统。该系统使用 3 次动态规划方法分割公式符号,并在此基础上使用层次结构方法分符号间的空间结构关系。在由 3268 个公式图像组成的测试集上取得了 78.2% 的 1 选公式分析正确率。实验结果验证了本文的数学公式识别系统的性能达到了初步应用的水平。下一步的工作是针对公式识别系统中出现的错误建立错误处理模型类型,减少错误的影响范围以避免出现关联性错误。

参 考 文 献

- [1] K-F Chan, D-Y Yeung. Mathematical expression recognition: A survey[J]. International Journal on Document Analysis and Recognition, 2000, 3(1): 3-15
- [2] Richard J Fateman, Taku Tokuyasu. Progress in recognizing typeset mathematics[C]. The SPIE, San Jose, CA, 1996
- [3] Hsi-Jian Lee, Jiumn-Shine Wang. Design of a mathematical expression recognition system[J]. Pattern Recognition Letters, 1995, 18(3): 289-298
- [4] Richard J Fateman, Taku Tokuyasu, Benjamin P Berman, et al. Optical character recognition and parsing of typeset mathematics[J]. Journal of Visual Communication and Image Representation, 1996, 7(1): 2-15
- [5] Masayuki Okamoto, B Miao. Recognition of mathematical expressions by using the layout structure of symbols[C]. In: Proc of the 1st Int'l Conf on Document Analysis and Recognition. Los Alamitos, CA: IEEE Computer Society Press, 1991. 242-250

- [6] Masayuki Okamoto, A Miyazawa. An experimental implementation of document recognition system for papers containing mathematical expressions [G]. In: Structured Document Image Analysis. Berlin: Springer-Verlag, 1992. 36-53
- [7] Hashim M Twaakyondo, Masayuki Okamoto. Structure analysis and recognition of mathematical expressions [C]. In: Proc of the 3rd Int'l Conf on Document Analysis and Recognition. IEEE Computer Society Press, 1995. 430-437
- [8] Masayuki Okamoto, Hiroki Imai, Kazuhiko Takagi. Performance evaluation of a robust method for mathematical expression recognition [C]. In: Proc of the 6th Int'l Conf on Document Analysis and Recognition. Los Alamitos, CA: IEEE Computer Society Press, 2001. 121-128
- [9] Jin Jianming. Research on typeset mathematical expression image processing: [Ph D Dissertation I D]. Tianjing: Nankai University, 2003 (in Chinese)
(靳简明. 数学公式图像理解研究 [博士论文 I D]. 天津: 南开大学, 2003)
- [10] Hao Yu, Ye Shiwei. Study of the counter propagation network based on the EM algorithm and its application [J]. Journal of Computer Research and Development, 2006, 43(5): 856-861 (in Chinese)
(郝玉, 叶世伟. 基于 EM 算法的对传网络学习与应用 [J]. 计算机研究与发展, 2006, 43(5): 856-861)
- [11] Biing-Hwang Juang, Wu Chou, Chin-Hui Lee. Minimum classification error rate methods for speech recognition [J]. IEEE Trans on Speech and Audio Processing, 1997, 5(3): 257-265
- [12] S F Chen, J Goodman. An empirical study of smoothing techniques for language modeling [J]. Computer Speech and Language, 1999, 13(10): 359-393



Guo Yusheng, born in 1979. Received his M. A 's. degrees in control science and engineering from Harbin Institute of Technology in 2004. Since 2004, he has been Ph. D. candidate in pattern recognition and intelligence system from the Institute of Automation, the Chinese Academy of Sciences, Beijing, China. His current research interests include character recognition, image processing, image understanding and machine learning.

郭育生, 1979 年生, 博士研究生, 主要研究方向为文字识别、图像理解、机器学习。



Huang Lei, born in 1977. He has been associate professor of the Institute of Automation, the Chinese Academy of Sciences since 2006. His main research interests are character recognition, image processing, image understanding and machine learning.

黄磊, 1977 年生, 副研究员, 主要研究方向为文字识别、图像理解、机器学习。



Liu Changping, born in 1965. Research professor and Ph. D. supervisor of the Institute of Automation, the Chinese Academy of Sciences. His main research interests include character recognition, image processing, image understanding and machine learning.

刘昌平, 1965 年生, 研究员, 博士生导师, 主要研究方向为文字识别、图像理解、机器学习。

Research Background

Optical character recognition (OCR) systems for mathematical documents which contain not only ordinary texts but also ME have been investigated. The development of such OCR provides many merits such as storage size reduction, search services and format conversion. Especially, the OCR for mathematical documents is indispensable on digitizing numerous historical mathematical documents for digital library. In this paper, a mathematical expression recognition system is proposed. During the symbol segmentation period, a three-stage segmentation method based on dynamic programming (DP) is proposed. During the structure analysis period, hierarchical structure is adopted to analyze the mathematical expression structure. Spatial relation model is used to calculate confidence of spatial relation between two symbols. The experiments were implemented on a database with 3268 mathematical expression images and the results show that the proposed system works well. Up to now, the system introduced has been integrated in a commercial OCR. Our work is supported by the National "863" High Technology Research and Development Program of China (2006AA01Z153).