联机手写化学公式识别与分析

杨巨峰，史广顺，王 恺
（南开大学信息技术科学学院，天津 300071）

摘 要：随着移动设备的日渐普及，联机手写输入方式为化学知识的使用和分享提供了可能，而化学公式的数字化处理也逐渐成为热点。为了有效进行联机手写化学公式的识别，通过整理 6 种化学公式中常见的符号位置关系，提出了一种用于联机手写化学公式识别与分析的方法，该方法在处理了断笔、粘连、连笔等书写异常情况后，完成了对化学公式的切分。识别时，先利用 SVM + HMM 的两级分类机制识别独立的化学符号；然后以公式的语义和语法规则协助理解用户的书写原意。实验证明，该方法应用于平板电脑，对于上述 3 个阶段的化学公式识别均取得了理想的结果，从而为联机手写化学公式重现和重用打下了基础。

关键词：联机手写化学公式；两级分类；结构分析
中图法分类号：TP391.4 **文献标志码**：A **文章编号**：1006-8961(2010)09-1291-08

Recognition and analysis of online handwritten chemical formulas

YANG Jufeng, SHI Guangshun, WANG Kai
(*College of Information Technical Science, Nankai University, Tianjin* 300071)

Abstract: With the development of mobile devices, the pen-based input mode has provided the possibility to use and share chemical knowledge easily. As an application of ubiquitous computing, research on online handwritten chemical formulas becomes a hot area. In this paper, we conclude common relations between chemical symbols and segment a chemical formula after linking the broken strokes. Then a two-level classifier is used to recognize the isolated symbols and the formula is understood with some chemical rules. The experimental results show that our method is robust and feasible when used on Tablet PCs.

Keywords: online handwritten chemical formulas; two-level classification; structure-based analysis

0 引 言

化学公式是表示化学反应规律的式子，是化学作用过程最重要的表现形式。和数学公式一样，它也拥有非常广泛的应用场景。为此，探索快速、高效的化学公式（结构）数字化方式就成为研究热点。相关研究始于 20 世纪 80 年代末，处理的对象包括印刷体化学文献文档、脱机手写化学文献文档和联机输入的化学公式 3 种。

Contreras 等人最早开始编制可以识别印刷体化学公式的程序[1]；随后 McDaniel 等人提出了多边形近似算法和检测虚线的算法[2-3]，并开发了第 1 款商业产品 Kekulé。同期，Ibison 等人实现了一个名为 CLiDE（化学文献数据提取）的原型系统[4]，从一幅原始文档图像中依次得到连续轮廓片段、连续线基元、虚线和字符，最终的解释结果存储为一张结构连接表。Casey 等人在总结前人工作的基础上，通过引入尺寸和空间特征来定位化学表达式，得到了理想的处理结果[5-6]。

针对脱机手写化学公式的识别问题，Ramel 等人将其分为文档全局感知和图形实体抽取两个阶段

基金项目：天津市自然科学基金项目（05YFJMJC01500）；中央高校基本科研业务费专项资金项目（65010201）。
收稿日期：2010-03-22；**改回日期**：2010-04-22

第一作者简介：杨巨峰（1980— ），男。讲师。2009 年于南开大学获得工学博士学位。主要研究方向包括人机交互、模式识别、人工智能等。E-mail: yangjufeng@nankai.edu.cn。

处理[7]。文献[7]方法包含了图像的结构化重现功能,并且赋予所有形状精确的描述,具体处理步骤是识别手写文本、定位多重连接、分析多边形及其相互关系,最后将获得的识别结果保存为向量形式。

随着平板电脑、手写板等硬件设备被越来越多的人接受,笔式人机交互技术已日趋成熟。一些研究者开始关注如何在这些移动设备上更方便地使用化学符号信息,化学公式的研究热点也转向了联机手写公式的识别领域。代表性的是 Tenneson 设计的一款用于平板电脑的化学教学软件 ChemPad[8],它可以逐笔识别手写字符的有机片段,并将其组合为 3 维形状。

国内专门针对化学公式识别的研究比较少。中国科学院软件所在笔式人机交互领域进行了扎实有效的工作,已经将手写字符识别技术成功应用于数学公式处理领域[9-10]。在此基础上,姜映映等人在 2006 年提出了一种手写化学公式的在线切分识别方法[11]。该方法在忽略异常符号、笔划跳跃和相连等复杂情况后,主要对切分算法进行了研究。其基本思想是:先对最近输入的 6 个笔划进行切分,然后从中选择最有可能的切分结果,已经被接受的笔划不再参与切分。在经过补充纠错、配平等后处理机制后,再引入语音识别技术构成一个多通道化学公式编辑器。

综合分析前期工作可以看出:以 3 种数据来源进行划分,其中针对印刷体化学公式的研究开展得最早、最充分,取得的成果也最多,相关技术已经比较成熟;脱机手写化学公式的应用场景很少,使得学者研究这一问题的热情不高;目前来看,联机手写化学公式的识别逐渐成为热点[12-13]。这是因为一方面越来越多的化学计算和表达工作需要借助计算机或其他电子设备完成,而且提高化学知识信息化、数字化水平的要求也日益迫切;另一方面,移动计算设备和相关支撑技术的发展确实为这一目标的实现提供了可能。

本文首先基于版面结构信息对化学公式进行切分,并综合运用垂直投影、连通域、字符宽度和部件结构等技术分析公式的组成关系,特别针对联机手写公式样本中常见的断笔、粘连、连笔情况进行处理;然后使用支持向量机(SVM)和隐马尔可夫模型(HMM)设计了一个两级分类器,用来对选定的 102 种化学符号进行分类和识别;最后根据化学公式的语法特点,按照“生成符→分隔符→文本区域→有机环结构”的顺序依次提取公式的主要成分。对于有机环结构通过分析其原子所在位置及原子间的键连接方式,以邻接矩阵的形式存储。依据本文提出的方法,一条手写化学公式就被逐步转换为内存中的一棵公式语法树,该结构可用于化学公式的存储、重现和检索。

1 公式版面结构分析

一般地,在进行联机手写公式处理流程中,需要先将对象切割成独立的符号再进行识别,研究者提出了许多成熟有效的方法[14-15]。化学公式是一种 2 维空间结构,无形中给符号切分造成了困难。本文综合使用多种结构分析技术来分析公式的结构,并在流程中添加了对断笔、粘连和连笔 3 种常见干扰的处理。

1.1 结构分析技术

1.1.1 投影

对图像点阵区进行 X 轴和 Y 轴方向上的投影,以得到横向和纵向的采样点统计直方图。字符点阵区域在直方图上呈现出波峰状,字符间的空隙呈现出波谷状。如果符号间没有发生粘连,则间隔区域在垂直投影方向上具有非常小的厚度,处理时可将局部极小投影值点作为候选切分点。该算法虽简单快速,但有如下不足之处:只适应很少粘连的情况、一个断裂的字符会被切分成几个字符,且不适应字符交叠的情况,所以本文仅使用该方法进行粗切分。

图 1 是利用投影法对一条化学公式进行切分的示意,由图 1 可以看到,只要阈值设定得当,至少可以保证不产生过切分的情况。

图 1 投影法粗切分公式

Fig. 1 Projection for segmentation

1.1.2 上下标

手写化学符号之间有几种频繁出现的位置关系。如果以一个符号为基准,另一个符号出现在它上标位置则表示这是一个离子表达式,而出现在它下标位置则表示该元素在物质中出现的次数。本文设计下面的算法,用来区分化学表达式中的上标、下标和水平 3 种位置关系。

图 2 表示一个手写硫离子字符,符号“S”和符号“2-”的外接矩形分别用($x_{1,1}$,$y_{1,1}$,$x_{1,2}$,$y_{1,2}$)、($x_{2,1}$,$y_{2,1}$,$x_{2,2}$,$y_{2,2}$)表示,且满足 $x_{i,1} < x_{i,2}$ 和 $y_{i,1} < y_{i,2}$,h_1 和 h_2 分别代表两个符号外接矩形的高,C_1 和 C_2 是两个符号外接矩形在竖直方向上的中点,B_1 和 B_2 分别代表两个符号重心,定义

$$\begin{cases} d = 0.7 \times y_{1,2} - y_{2,2} + 0.3 \times y_{1,1} \\ T = 1000 \times d/h_1 \\ B = 1000 \times (B_1 - B_2)/h_1 \end{cases} \tag{1}$$

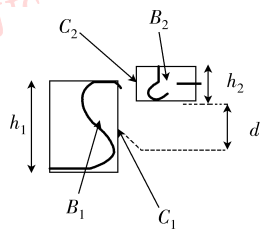

图 2　符号“S”和“2-”之间的位置关系

Fig. 2　Spatial relation between “S” and “2-”

本文对文献[16]的方法做了改进,引入特征对(T, B)用来区分上标、下标、水平 3 种关系。图 3 是在 65 个化学符号对样本上所做的统计,除有个别下标被误识为水平关系外,超过 95% 的符号关系都能正确识别,能满足本文实验需要。

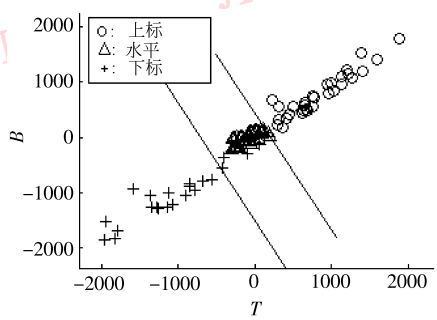

图 3　(T,B)区分位置关系结果

Fig. 3　Determine spatial relations using (T,B)

1.1.3　部件

以上下标分析为基础,本文进一步提出了一种基于部件的公式切分技术。一条化学公式经过投影法切分后即产生若干字符体(可能是几个字符的集合),称之为部件,部件由若干个部件元组成。其中,一个部件元的位置信息包括宽度、高度、左上角坐标、右下角坐标、中心点坐标。把部件元按部件元中心的 X 轴坐标排序来分析部件元之间的关系。当两个相邻部件元具有左右、右下关系时,则不可以合并,否则将其合并。最后所得到的一个部件元就是一个化学符号,所以一个部件的部件元个数就是该部件被分割成的字符数。图 4 显示了联机手写化学公式中常见的 6 种部件关系。

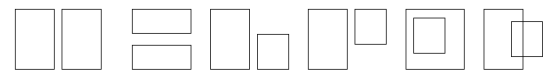

图 4　部件关系

Fig. 4　Relations between two components

1.2　处理异常输入

尽管使用上述方法对化学公式的版面结构进行分析,在书写规范的公式样本上可以取得很好的效果,然而由于用户书写的随意性和硬件设备的采集效果不佳等原因,会使公式样本中经常出现断笔、粘连、连笔等特殊情况。对这些异常情况的处理直接关系到整条公式结构分析的结果,同时也决定了切分方法能否做到与用户无关。

1.2.1　处理断笔

在联机手写化学公式中,由于书写设备或人为的原因而经常出现符号的某一笔画断裂为两个或多个片段的情况,这就是断笔。解决联机识别问题时,因为笔划是一个重要的信息单元,所以断笔的出现会对符号特征的提取产生严重影响,进而降低识别的准确率。此外,如果在切分阶段将本来属于同一符号的断裂笔划,划分到不同单元中去,也会干扰整条化学公式的分析和理解。图 5 是几种断笔的情况。

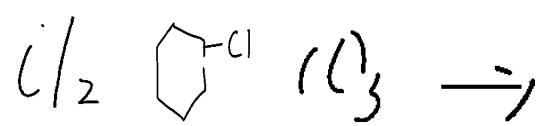

图 5　断笔符号

Fig. 5　Broken symbols

本文结合笔画方向特征及书写时间特征来处理化学公式中的断笔,显然判断的基础是这两个笔划相邻,步骤如下:

1) 定位孤立点,找到后再设法将其与之前或之后的一个笔画合并;

2) 对时间间隔信息进行判断,计算两个相邻笔划的时间间隔是否小于整条公式的笔划平均时间间

隔,如果是,则认为这两个相邻笔划可能是一笔划断开的情况;

3)在可疑区域的前一笔划尾部取两点 A 和 B,后一笔划首部取两点 C 和 D,分别计算 AB 与 BC 和 BC 与 CD 的角度变化。如果小于某个阈值,则认为二者的书写方向一致,连接它们。

1.2.2 处理粘连

粘连是指空间相邻的两个符号由于书写时过于靠近而贴在一起。粘连涉及的字符分属不同笔划,一般不存在额外的子笔划连接。图 6 是一些粘连在一起的符号。处理粘连的步骤如下:

1)在各连通域中,检测是否存在有机环结构,如果存在,则将其与外围粘连的化学键切分开;

2)根据位置关系和粘连程度定位右上粘连和右下粘连的情况;

3)判断焦点区域是否与预先设定的某些粘连模板匹配,如果是,则切分所涉及到的符号。当扫描全部连通域,并重复执行上述步骤之后,则所有的粘连区域都被处理。

图 6 粘连符号

Fig. 6 Connected symbols

1.2.3 处理连笔

与粘连问题不同,连笔是多个字符由一笔写成的情况。图 7 显示了化学公式中一些连笔写成的符号。

图 7 连笔符号

Fig. 7 One-stroke symbols

虽然联机手写化学公式中产生连笔的原因很多,但是通过观察和分析连笔的样本可以总结出一定规律,即绝大多数连笔情况都具有以下两个特征:

1)连笔的笔划长度必然大于整条化学公式中所有笔划长度的平均值。这里的笔划长度是指笔划上所有采样点按书写顺序计算得到的各点间欧几里德距离的总和,即

$$l_{\text{Stroke}} = \sum_{i=0}^{N-1} D_{i,i+1} \tag{2}$$

其中 N 是手写笔划上采样点的个数,$D_{i,i+1}$ 是第 i 个点与其相邻点之间的欧氏距离。

2)无机化学符号中大多数连笔的书写走势是先由上到下,然后由下到上,最后由上再到下。如图 8 所示,在书写时,连续发生两个方向转换后,黑圆点处就是可能的连笔切分点。

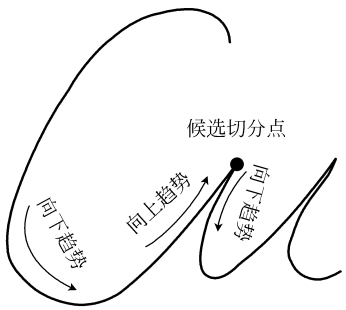

图 8 连笔的一般走势

Fig. 8 A general tendency of connected strokes

本文提出的连笔处理方法如下:

1)首先对输入的联机手写化学公式进行预处理,在连笔判断和切分前进行一些常规的预处理操作可以有效减少误切分和过切分;

2)检测有机环结构是否存在,如果有,则依次将其切分出来,同时,当有其他笔划的外接矩形包含在有机环的外接矩形里时,则把它们合并在一起,作为一个完整的有机环符号;

3)根据书写时间以及基本走势,先定位可疑区域,然后通过比对模板补充其他候选切分点;

4)根据定位到的切分点,分析连接符号的位置关系和连接方式。如果属于共享点连接方式,则直接在切分点处切开;如果属于通过一段额外的子笔划连接起来的符号,则去除多余的子笔划。

1.3 切分实验结果

对于联机手写化学公式结构分析中遇到的断笔、粘连和连笔 3 种问题,本文提出了一种基于结构特征的解决办法。切分实验是基于惠普平板电脑采集的 1200 条公式样本进行的,采集工作分 6 次进行,每次由 10 名化学专业及非化学专业学生在无监督的状态下任意书写 20 条联机化学公式。

这些样本的统计结果为:包含断笔的化学公式 217 条、其中断笔区域 281 个;包含粘连符号的化学公式 261 条,其中粘连区域 330 处;包含连笔情况的

化学公式 257 条,其中连笔笔划 382 个。利用本文方法进行处理后,实验结果如表 1 所示。

表 1　切分实验结果

Tab. 1　Segmentation result

指标	断笔	粘连	连笔
问题区域数	281	330	382
正确操作数	268	306	355
正确操作率/%	95.4	92.7	92.9
误操作数	6	2	3
误操作率/%	2.1	0.6	0.8

通过分析上述切分结果及发生错误的样本发现:脏点的存在会对孤立点的寻找造成干扰,而且算法认为,断笔和连笔的时间间隔应该分别小于和大于两个正常笔划的时间间隔的假设在某些特殊样本上不成立。不过出现问题的符号比较集中,容易引入模板匹配的方法解决。同时本文算法在实验样本集上的过切分率非常低,能够保证在后续应用中不会将一个符号错误地切分为两个部分。

2　手写化学符号识别

2.1　识别对象和方法框架

联机手写化学符号是出现在化学公式中的具有独立化学意义的笔划或笔划集合。识别化学符号既是手写化学公式向其计算机表达转化的基础,同时也为理解化学公式的语法和语义提供了必要的支撑。本文处理的符号包含英文字母、阿拉伯数字、化学操作符和各种有机环结构共 102 种,方法框架如图 9 所示。

使用基于 SVM 和 HMM 的两级分类方法来训练和识别化学符号。识别前首先将采集到的联机手写化学符号实验样本集按照 3∶1 的比例分为训练集和测试集两部分;然后在训练集上训练环/非环两类模型和具体的符号模型,并将得到的模型参数加载到识别流程入口。识别时,通过计算 SVM 特征,确定输入的是一个有机环符号,还是普通符号,进而在对应的子类中计算属于各个 HMM 模型的概率,与概率最大的模型对应的符号即为识别结果。

2.2　环/非环粗分类

第 1 级分类是使用 SVM 划分环和非环。SVM 是统计学习理论的重要组成部分,与传统方法比较,它具有坚实的理论基础和较好的推广能力、优秀的非线性处理能力和高维处理能力。对于两类模式分类问题,它不但有着优美而直观的表达,而且处理效率也在同类方法中领先。

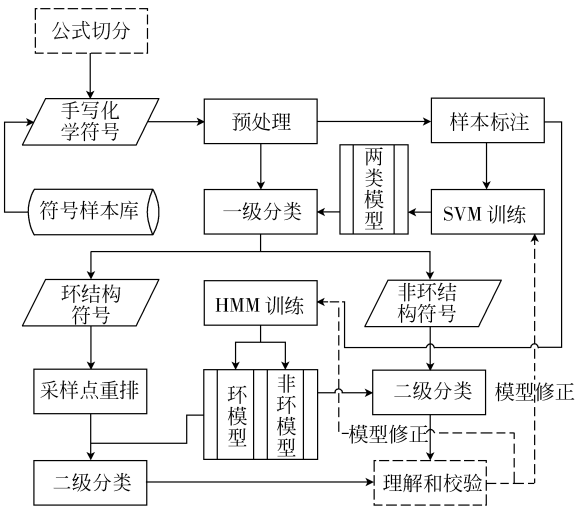

图 9　化学符号识别模型

Fig. 9　Recognition model of chemical symbols

通过对环与非环样本的观察发现:两类样本在采样点的坐标分布上具有明显的差别。多数非环样本的采样点分布较均匀,而对于环样本,采样点主要分布在样本的边界,在样本中心分布较少,甚至没有。对环与非环样本具有较好区分能力的另一种特征是外围轮廓特征,它是一种能够刻画给定图像外部边界形状的全局特征。考虑到环符号与非环符号在外围轮廓上的差别,可按先后顺序从样本图像的左、下、右、上 4 边分别向右、上、左、下 4 个方向扫描,直至扫描线遇到笔划或中轴,记下各自扫描线经过的距离即为该样本的外围轮廓特征。

2.3　符号细分类

第 2 级分类是使用 HMM 区分符号的细致类别。由 HMM 的定义可知,它由处理语音信号的需求而产生,对于具有明显时序性质的信息有良好的建模能力。联机手写的化学符号由一系列离散的采样点组成,由于这些点正是按照书写的时间顺序排列起来的,所以联机手写化学符号的时序特性决定了它适合于使用 HMM 的方法进行处理[17-18]。

从化学符号的第 1 个采样点开始逐点记录其特征,并通过选取 11 维局部特征来涵盖特征点所处区域完整的位置信息和方向信息。

对有机环符号需要进行一些额外的处理。这是因为书写复杂结构符号时,人们往往先将其拆分成

一些基元(基本笔划)的组合,然后按照一定的次序逐个书写,最终形成整个符号。就不同的书写者而言,同一个有机符号可能划分为不同的基元,而且即使基元组合相同,其先后书写次序也可能不相同。甚至一个人在两次书写同一个符号时也可能有不同的落笔顺序。

为了减少上述事实对 HMM 建模和识别的影响,在抽取有机环结构的 11 维局部特征前,应首先对其采样点集合进行重新排列,以使得同一符号的不同样本尽量具有类似的时序特性。重排过程如图 10 所示。

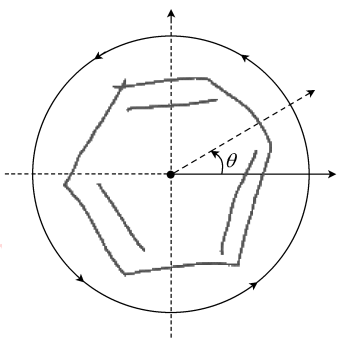

图 10 有机环重排算法

Fig. 10 Sort points in a circle

重排算法如下:

1) 计算有机符号样本的质心;

2) 将由质心向外引出的一条射线记为扫描线,而将扫描线与 X 轴正向的夹角(方向角)记为 θ,将 θ 初始值设为 0;

3) 遍历有机符号的点序列,计算每个点到扫描线的距离 d,如果 d 小于一个预先设定的阈值 F,则将此点存入重排点队列 ***List***,否则不予处理;

4) 沿逆时针方向将扫描线旋转一个角度 $\Delta\theta$,此时扫描线的方向角为 $\theta+\Delta\theta$;返回到上一步,直至重排完毕。

2.4 识别实验结果

对于基于 SVM 的有机环/非有机环两类分类问题本文利用如下 3 种核函数进行了实验:

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\left(\frac{(\boldsymbol{x}\cdot\boldsymbol{x}_i)}{256}\right)^d,d=1,2,\cdots,6 \quad (3)$$

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\{-\gamma\|\boldsymbol{x}-\boldsymbol{x}_i\|^2\} \quad (4)$$

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\tan\left(\frac{(\boldsymbol{x}\cdot\boldsymbol{x}_i)}{128}-1\right) \quad (5)$$

实验采用的数据集包括 9 180 个训练样本和 3 264个测试样本。训练样本事先标注了所属的类别,而测试样本则用来检验各种参数配置下的分类效果。具体实验结果如表 2 所示。

表 2 SVM 分类结果

Tab. 2 SVM recognition result

d	γ	C	识别率/%	
			训练集	测试集
3	—	—	98.67	98.96
—	2	—	100	**99.82**
—	—	2^{15}	99.73	99.51

由表 2 可得到如下结论:基于径向基内积函数的 SVM 在相同的测试集上,能获得最高的识别率,此时的最佳参数组合为:惩罚因子 $C=2^{11}$,径向基核函数的参数 $\gamma=2$。利用该参数配置测试得到的识别率是 99.82%,即在 1 000 个任意书写的联机手写化学符号中,仅有不到两个符号被划分到错误的环/非环类别。

将样本集进一步细分为有机环结构的集合和普通无机符号集合,组合不同状态数与高斯数,在测试集上得到的符号分类实验结果如表 3 所示。

表 3 HMM 分类结果

Tab. 3 HMM recognition result

高斯数	状态数	识别率/%			
		环 Top1	环 Top3	非环 Top1	非环 Top3
6	4	96.0	99.1	87.9	98.4
6	6	96.8	99.3	88.8	98.3
6	8	97.0	99.1	89.0	98.5
9	4	96.5	99.1	87.9	98.3
9	6	96.7	99.3	**89.5**	98.4
9	8	**98.5**	99.9	89.3	98.7

由表 3 可得到如下结论:对于有机环结构,Top-1 上的最高识别率为 98.5%,对应这一结果的 HMM 参数如下:8 状态,每个状态 9 个混合高斯,此时正确结果出现在前 3 个候选项中的概率接近 100%;对于无机符号,首候选项的最高识别率为 89.5%,对应的 HMM 参数如下:6 状态、每个状态 9 个混合高斯。在这种参数配置下,前 3 个候选项的识别率达到了 98.4%,若辅以适当的筛选和修正规则后即可以满足实用要求。

3 语法规则指导的公式理解

联机手写化学公式处理的一个重要任务是:理

解公式代表的实际化学含义,回答哪些物质参与了反应、生成了哪些物质、生成物以何种状态存在、各物质质量符合何种关系等问题。

一般而言,化学公式是一个水平 1 维层面上的结构,即是反应物、生成物、化学反应符号在水平方向的组合结果。化学公式理解的困难,主要集中在物质的显示形式上,因为这是一个 2 维层面的结构。以一个随意书写的有机结构为例,该物质含有如图 11 所示的局部结构。

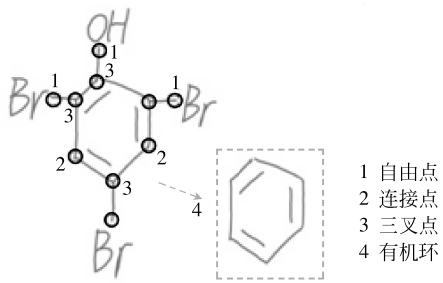

图 11　化学表达式中的局部结构

Fig. 11　Internal features

一般地,有机环、外接键和相关字符群组合的版式结构符合如图 12 所示的规律。

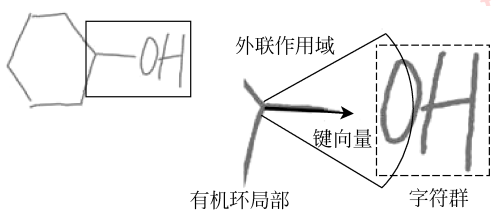

图 12　有机环外联结构

Fig. 12　External features

如果认为一个连接环结构和字符群的化学键派生自前者,则字符群所处的矩形必然位于该化学键向量的有限辐射域上。这一特性既可以用于环结构外围字符群的检测,也能指导有机表达式的语义校验。

可将化学公式抽象定义为一个六元组,形如 $\boldsymbol{G}=(\boldsymbol{S}, L, T, Gr, Sy, Se)$,其中 $\boldsymbol{G}$ 代表化学公式的语法结构,$\boldsymbol{S}$ 是符号集合,L 是版面信息,T 记录时序信息,Gr 是语法规则,Sy 是句法规则,Se 代表语义规则。

一条联机手写化学公式依次提取的生成符、分隔符、文本区域如图 13 所示,分析其环状结构,最终即得到一棵公式语法树。

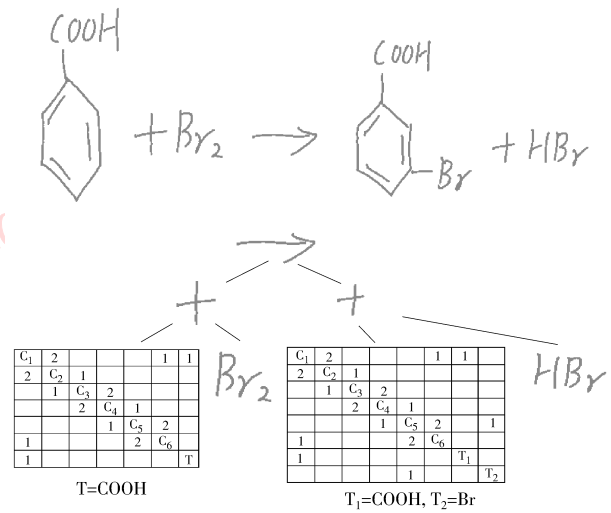

图 13　公式理解

Fig. 13　Formula understanding

4　结　论

本文针对联机手写化学公式处理的切分、符号识别和公式理解 3 个关键步骤展开研究,构建了比较完整的化学公式版面结构分析流程,由于考虑了各种特殊情况,因此本文方法可在无约束公式样本上应用。符号识别部分粗分类错误率可控制在 0.2% 以下,基于 HMM 的细分类准确率超过 90%。以化学公式的形式化描述,特别是以语法规则为基础设计了一种公式语法树结构,以便将公式理解转化为寻找合适的基元填充语法树节点的问题。

联机手写化学公式识别与分析为将笔式人机交互理论应用于化学领域提供了技术基础。在后续研究中,还应该在复杂结构处理、公式重现和重用等方面进行深入探索。

参考文献(References)

[1] Contreras M, Allendes C, Alvarez L, et al. Computational perception and recognition of digitized molecular structures [J]. Journal of Chemical Information and Computer Sciences, 1990, 30(3): 290-302.

[2] McDaniel J, Balmuth J. Kekulé: OCR—optical chemical (structure) recognition [J]. Journal of Chemical Information and Computer Sciences, 1992, 32(4): 373-378.

[3] McDaniel J, Balmuth J. OCR of chemical structure diagrams [C]//Proceedings of Computerized Chemical Data Standards: Databases, Data Interchange, and Information Systems. Philadelphia, Penn., US: American Society for Testing and Materials, 1994: 3-15.

[4] Ibison P, Jacquot M, Kam F, et al. Chemical literature data extraction: the CLiDE project [J]. Journal of Chemical Information and Computer Sciences, 1993, 33(3): 338-344.

[5] Casey R, Boyer S, Healey P, et al. Optical recognition of chemical graphics [C]//Proceedings of the 2nd International Conference on Document Analysis and Recognition (ICDAR93). Los Alamitos, CA, USA: IEEE Computer Society, 1993: 627-631.

[6] Boyer S, Casey R, Miller A, et al. Apparatus and Method for Parsing a Chemical String: United States, 5345516[P]. 1994-9-6.

[7] Ramel J, Boissier G, Emptoz H. Automatic reading of handwritten chemical formulas from a structural representation of the image[C]// Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR99). Washington, DC, USA: IEEE Computer Society, 1999: 83-86.

[8] Tenneson D. ChemPad: A Pedagogical Tool for Exploring Handwritten Organic Molecules [R]. No. RI 02912, Brown University, Providence, RI, USA: 2005.

[9] Feng Haibo, Tian Feng, Luan Shangmin, et al. Application of interactive techniques to handwritten mathematics editing [J]. Journal of Computer-aided Design & Computer Graphics, 2003, 15(11): 1437-1442. [冯海波,田丰,栾尚敏,等. 交互技术在手写公式编辑中的应用[J]. 计算机辅助设计与图形学学报, 2003, 15(11): 1437-1442.]

[10] Feng Haibo, Li Zhaoyang, Dai Guozhong. Gesture-based handwriting mathematics editing system [J]. Computer Engineering and Applications, 2003, 39(35): 97-99. [冯海波,李昭阳,戴国忠. 基于手势的手写公式编辑系统[J]. 计算机工程与应用, 2003, 39(35): 97-99.]

[11] Jiang Yingying, Wang Xugang, Ao Xiang, et al. Online recognition of handwritten chemical formula[C]//Proceedings of the 2nd Joint Conference on Harmonious Human Machine Environment (HHME06). Beijing, China: Tsinghua University Press, 2006: 111-116. [姜映映,王绪刚,敖翔,等. 手写化学公式的在线切分识别[C]// 第 2 届和谐人机环境联合学术会议论文集, 北京:清华大学出版社, 2006: 111-116.]

[12] Yang Jufeng, Shi Guangshun, Wang Kai, et al. A study of on-line handwritten chemical expressions recognition [C]// Proceedings of the 19th International Conference on Pattern Recognition (ICPR08). Piscataway, NJ, USA: IEEE Computer Society, 2008: 1-4.

[13] Zhang Yang, Shi Guangshun, Yang Jufeng. HMM-based online recognition of handwritten chemical symbols[C]//Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR09). Piscataway, NJ, USA: IEEE Computer Society, 2009: 1255-1259.

[14] Ding Jie, Lou Zhen, Yang Jingyu. Segmentation of numeral strings using stroke grouping [J]. Journal of Image and Graphics, 2009, 14(8): 1609-1614. [丁杰,娄震,杨静宇. 基于笔划组合的手写数字切分[J]. 中国图象图形学报, 2009, 14(8): 1609-1614.]

[15] Liwicki M, Scherz M, Bunke H. Word extraction from on-line handwritten text lines[C]//Proceedings of the 18th International Conference on Pattern Recognition (ICPR06). Piscataway, NJ, USA: IEEE Computer Society, 2006: 929-933.

[16] Jiang Hongying, Jin Jianming, Wang Qingren. Determine superscript/subscript relations in typeset mathematical expressions based on statistic features [J]. Computer Engineering and Applications, 2003, 39(28): 75-78. [江红英,靳简明,王庆人. 基于统计特征的印刷体数学公式上/下标关系判别[J]. 计算机工程与应用, 2003, 39(28): 75-78.]

[17] Artieres T, Marukatat S, Gallinari P. Online handwritten shape recognition using segmental Hidden Markov Models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 205-217.

[18] Hu Jianying, Brown M, Turin W. HMM based on-line handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(10): 1039-1045.