2009 10th International Conference on Document Analysis and Recognition

A Unified Framework for Recognizing Handwritten Chemical Expressions

Ming Chang, Shi Han, Dongmei Zhang
Microsoft Research Asia
{mingch, shihan, dongmeiz}@microsoft.com

Abstract

Chemical expressions have more variant structures in 2-D space than that in math equations. In this paper we propose a unified framework for recognizing handwritten chemical expressions including both inorganic and organic expressions. A set of novel statistical algorithms is presented in two key components of this framework: symbol grouping and structure analysis. Non-symbol modeling and inter-group modeling are proposed to achieve better grouping result, and bond modeling is proposed to group the special bond symbols in the unified framework. A graph-based representation (CESG) is defined for representing generic chemical expressions, and the structure analysis problem is formulated as a search problem for CESG over a weighted direction graph. Experiments on a database of more than 35,000 expressions were conducted and results are presented.

1. Introduction

Handwritten scientific expression recognition is one enabling technology to achieve natural user experience in human computer interaction especially in the education domain. While some promising results[1][3] [10][11] were reported and a few commercial software products[4][5] were released on recognizing handwritten math expressions in the past few years, the research on recognizing chemical expressions was much less active.

Chemical expressions include formula and equations from both inorganic and organic chemistry. While the inorganic expressions have strong structural resemblance to math expressions, diagram like organic formula are much more different and complex due to various bonds as illustrated in Figure 1. These characteristics present challenges not only on recognizing organic expressions, but also on designing a unified solution for recognizing both inorganic and organic expressions.

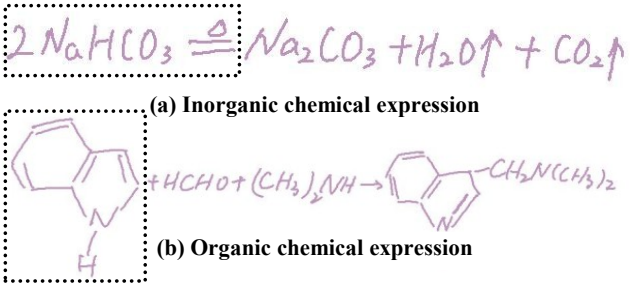


Figure 1: Examples for chemical expressions

Little work on handwritten chemical expression recognition was reported previously. Neural network was used in [8] to recognize chemical heterocyclic rings. A system was presented in [9] to recognize handwritten organic chemical compounds. SVM was used for symbol grouping and recognition, and local spatial context and a set of domain-specific constraints were used for structure interpretation. A system for auto generation of 3D views of handwritten organic molecules was proposed in [12] with the focus on 3D structure generation rather than recognition. Strong assumptions such as every symbol written in one stroke were needed for the recognition algorithm. [13] presented a system to recognize handwritten chemical formula image from a structural representation. [14] proposed a two-level algorithm to recognize handwritten chemical expressions. A useful set of strokes, structural and syntactic features are adopted for segmentation and recognition while some organic loop symbols are pre-defined. It didn't mention how to handle chemical compounds with arbitrary structure.

We propose a unified framework for recognizing both inorganic and organic expressions. The framework consists of three components – symbol grouping, structure analysis and semantic verification.

Given a series of ink strokes, symbol grouping segments the strokes into groups that constitute symbol candidates. In addition to modeling common chemical symbols, we create models for non-symbols and bond symbols in symbol grouping in order to reduce the rate of mis-grouping. Non-symbol and bond modeling also

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.64

enable symbol grouping to be conducted in a consistent manner for both inorganic and organic formula.

Once the input ink strokes are grouped into potential symbols, structure analysis is conducted to determine the structural relationship among symbols including bonds. We formulate structure analysis into a graph search problem in which based on defined criterion, the correct expression structure is the optimal graph of the weighted direction graph representing symbol relationship. Semantic verification leverages domain knowledge to refine the recognition candidates of structure analysis and produce the final result.

It should be noted that both spatial and contextual information is statistically incorporated into the proposed framework and delayed decision making is used at every stage of the statistical framework in order to search for the optimal recognition result using information from all components.

The technical contributions of our work are in three aspects. First of all, this is the first time a unified framework for recognizing both inorganic and organic expressions is proposed. Second of all, we propose a systematic symbol grouping algorithm that handles common symbols and bonds in a consistent manner. Third of all, we propose a Chemical Expression Structure Graph (CESG) to represent both inorganic and organic expressions as well as a weighted direction graph representation for searching CESG to identify the structures of chemical expressions.

The rest of the paper is organized as follows. The details of symbol grouping and structure analysis are discussed in Section 2 and 3, respectively, followed by discussion on semantic verification in Section 4. Experimental results are presented in Section 5. We conclude the paper in Section 6.

2. Symbol Grouping

Given a stroke sequence $(o_1, o_2, \ldots, o_N)$ of N strokes, the goal of symbol grouping is to find the optimal symbol sequence $(G_1, G_2, \ldots, G_n)$ with corresponding boundaries in the maximum likelihood sense as shown in Eq (1)

$$G_{\max} = \max_{(G_1, G_2, \ldots G_n)} P((G_1, G_2, \ldots G_n) \mid (o_1, o_2, \ldots o_N)) = \max_{(G_1, G_2, \ldots G_n)} \prod_{i=1}^{n} P(S \mid G_i) P(G_i \mid G_{i-1}) \quad (1)$$

where $P(S|G_i)$ is the probability of group G_i forming a symbol S and $P(G_i|G_{i-1})$ is the probability that group G_i is the successor of group G_{i-1}.

Dynamic programming is usually used to optimize the overall symbol recognition results across the entire sequence. The key to achieving high grouping accuracy lies in the quality of individual symbol recognition result and the utilization of other information that helps reduce the underlying grouping confusion. Most approaches often use symbol recognition results and positional stroke relation as clues. Spatial grammars on structure symbols are defined and used in symbol grouping in [10]. Here we present novel techniques that not only improve symbol recognition confidence but also incorporate spatial and contextual information into symbol grouping.

2.1. Non-symbol modeling

Different from individual symbol recognition, recognizing a symbol in a stroke sequence requires that a symbol hypothesis be rejected, i.e. extremely low recognition score $P(S|G_i)$ is provided, when the underlying stroke sequence does not constitute a valid symbol, e.g., when over-grouping occurs. Otherwise symbol boundaries cannot be determined correctly. This means that the model space spanned by the valid symbol models is incomplete and it must be supplemented with non-symbol models.

In our implementation, we build non-symbol models iteratively using training data. The initial non-symbol model is built using the incorrect grouping results obtained without non-symbol model. Then non-symbol model is added to conduct symbol grouping on the training data again and better grouping results are obtained. The initial non-symbol model is improved using the incorrect grouping results from the second round. As the process is repeated, both the non-symbol model and the symbol grouping result are improved.

2.2. Inter-group modeling

The rich spatial and contextual information between neighboring symbol groups can be captured by inter-group modeling (Eq(2)) which is effective in reducing under-grouping errors.

$$P(G_i \mid G_{i-1}) = \sum_{S_1, S_2, R_j} P(G_{i-1} \mid S_1) P(G_i \mid S_2) P(G_{i-1}, G_i, R_j) P(S_2 \mid S_1, R_j) \quad (2)$$

In Eq(2), S_1 is the symbol candidate of stroke group G_{i-1}, S_2 is the candidate of the successor group G_i, and R_j is the possible spatial relationship between G_{i-1} and G_i. $P(S_2|S_1,R_j)$ represents the occurring probability of S_2 given the preceding symbol S_1 and relationship R_j. $P(G_{i-1}|G_i, R_j)$ represents the probability of G_{i-1} and G_i with relationship R_j. As shown in Figure 2, nine spatial relations are defined for R_j based on the characteristic of chemical formula. $P(G_{i-1}|G_i, R_j)$ is calculated using Gaussian Mixture Model.

2.3. Bond modeling

Bonds are special symbols in organic formula. There are three types of bonds – single bond, double bond and triple bond (Figure 3a). Same as other chemical symbols, these bonds need to be recognized in symbol grouping in order for structures containing these bonds to be identified correctly. Handwritten bonds are difficult to recognize because bonds can be written using various number of strokes and late strokes are often present in bonds (Figure 3b).

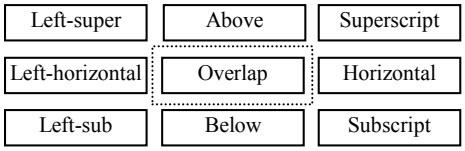
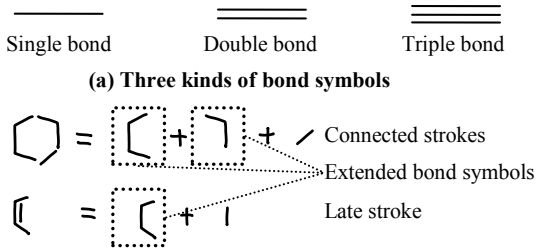


Figure 2: Spatial relations for inter-group modeling



(a) Three kinds of bond symbols

(b) Writing variance of bond symbols

Figure 3: Bond symbol

Unlike the heuristic based approaches, we propose a learning based algorithm to overcome the above challenges. First of all, an extended bond symbol is defined to represent all types of bonds that are written in one stroke (Figure 3b). The models of extended bond symbol are built using the training data set. The extended bond symbol is treated the same way as other symbols in symbol grouping, i.e. candidate groups of extended bond symbol will be present if the underlying expression consists of organic formula. At the second step, each extended bond symbol is split into single bond using a Curvature Scale Space (CSS) [6] based corner detection algorithm. Then neural network is used at the last step to verify the resultant bonds from the second step and group the valid ones into single, double and triple bonds.

Using the above bond modeling algorithm, symbol grouping of inorganic and organic expressions is unified in the same recognition framework.

3. Structure Analysis

The output of symbol grouping is a sequence of stroke groups each having *n* candidate symbols. The goal of structure analysis is to identify the structural relationship among the stroke groups and determine the recognized symbol for each stroke group. In this section, we first propose a representation named Chemical Expression Structure Graph (CESG) to represent both inorganic and organic expressions and then formulate structure analysis as a search problem for CESG over a weighted direction graph constructed using the symbol grouping result.

3.1. Chemical Expression Structure Graph

Chemical expressions have two-dimensional structure and this structure is further complicated by the enormous ways that atoms and molecules are connected by bonds in organic chemistry (Figure 1b).

A Chemical Expression Structure Graph $G_{CES}(V, E)$ is a directed graph in which $V = \{v, v \in V_{chem}\}$ is the set of chemical symbols in a given chemical expression. Here V_{chem} is the set of all chemical symbols with two subsets as defined in (3).

$$V_{chem} = V_{bond} \cup V_{nonobond}, V_{bond} = \{-, =, \equiv\} \quad (3)$$

$E = \{e, e \in E_{connection}\}$ is the set of edges connecting the symbol vertices according to the expression structure. As defined in (4), there are four types of connections between symbols.

$$E_{connection} = \{E_c, E_{nc}, E_{peer}, E_{sub}\} \quad (4)$$

E_c and E_{nc} can only connect bond symbols. E_c means that two bond symbols are indirectly connected via an atom and E_{nc} means that the connection is direct with atom omitted, e.g. Carbon. Any two non-bond symbols as well as one bond symbol and one non-bond symbol can be connected by either E_{peer} or E_{sub}. E_{peer} means that two connected symbols are equal in terms of spatial position. Any vertex has at most one incoming and outgoing E_{peer}, respectively. E_{sub} means that one symbol is spatially subordinate to the other. It always originates from non-bond symbol; and at most one E_{sub} can originate from a non-bond symbol. Figure 4 shows examples of CESG for parts of expressions in Figure 1.

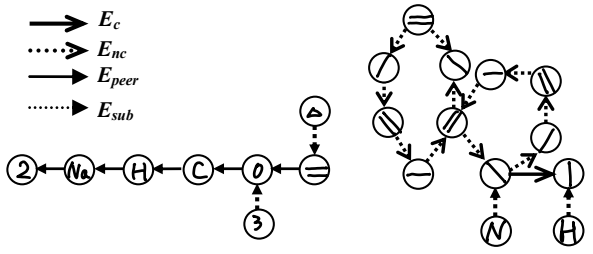


Figure 4: Examples of CSEG

3.2. Constructing Weighted Direction Graph

We now construct a weighted direction graph G_{WD} in order to conduct global and statistical search for the optimal CESG. This global approach can avoid unrecoverable errors caused by local optimization. A similar approach was presented in [2] for analyzing math expressions, but little discussion was found on calculating edge score in the graph to be searched.

The candidate symbols generated by symbol grouping are the vertices of G_{WD}. The same types of edges are defined as in (4). Possible edges between any pair of vertices are constructed and a weight representing the connection strength is assigned to each edge. The edge weight is calculated in (5) by incorporating the symbol recognition results, the spatial relationship between symbols and the contextual relationship of the two symbols.

$$E_{connection}(S_a,S_b,R) = P^{\alpha}(O_a \mid S_a)P^{\alpha}(O_b \mid S_b) \\ P^{\beta}(O_a,O_b \mid R,S_a)P^{\gamma}(S_b \mid S_a,R) \quad (5)$$

In (5), S_a and S_b are symbol candidates; R is the edge type representing the relationship between S_a and S_b. Generated by the symbol grouping component, $P(O_a|S_a)$ and $P(O_b|S_b)$ are the probabilities that two symbols are recognized as S_a and S_b, respectively. Calculated using the spatial model and the contextual model respectively, $P(O_a,O_b|R,S_a)$ and $P(S_b|S_a,R)$ represent the probability that symbol S_b is subordinate to symbol S_a under relationship R. Parameters α, β and γ are used to combine the above probability values into one edge score $E_{connection}(S_a,S_b,R)$.

We now use the edge type E_{sub} to explain how to calculate $P(O_a,O_b|R,S_a)$. We define a rectangular control region (Figure 5) for S_a and examine the spatial distance from S_b to the control region. The shorter the distance is, the more likely S_b is subordinate to S_a. Since O_b is the stroke group of S_b, the sample points of O_b are used to calculate this spatial distance as the probability of O_b in the control region. For a sample point (x_i, y_i), this probability is calculated using (6) where $O_{R(x_i)}$ and $O_{R(y_i)}$ are the horizontal and vertical offsite to the control region, respectively;

$$P((x_i,y_i) \mid R,S_a) = \frac{1}{1+(\frac{|O_{R(x_i)}|}{x_0})^{\gamma x}} * \frac{1}{1+(\frac{|O_{R(y_i)}|}{y_0})^{\gamma y}} \quad (6)$$

$(x_0, y_0, \gamma_x, \gamma_y)$ are parameters of the distribution model. $P(O_a,O_b|R,S_a)$ is then calculated by averaging $P((x_i,y_i)|R, S_a)$ for all the sample points of O_b.

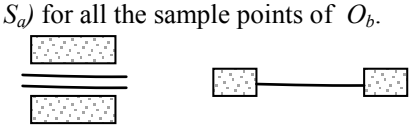


Figure 5: Rectangular control regions for reaction condition and bond

$P(O_a,O_b|R,S_a)$ for the other edge types are calculated using similar approaches. The parameters for control regions and distribution models are obtained via training on a fairly large data set.

There is strong contextual information between symbols due to the domain characteristics of chemistry. For example, the co-occurrence of atoms Carbon C and Oxygen O and the co-occurrence of single bond – and Carbon C. Using a large training set, this kind of contextual information can be modeled and used to calculate the last item $P(S_b |R, S_a)$ in (5) as shown in (7)

$$P(S_b \mid S_a,R) = \frac{P(R,S_a,S_b)}{P(R,S_a)} = \frac{C(R,S_a,S_b)}{C(R,S_a)} \quad (7)$$

Where $C(R, S_a, S_b)$ is the number of co-occurrences of symbol S_a and S_b with relationship R; and $C(R,S_a)$ is the number of occurrences of S_a having relationship R with any other symbol. Smoothing technique is used when there are not enough data for some symbols in the training data set.

3.3. Searching Optimal CESG

After the weighted direction graph G_{WD} is constructed, the structure of the underlying chemical expression can be obtained by searching for the optimal CESG on G_{WD} using the criterion defined as:

$$G^{CES}_{\max} = \max_{M_n,S_i^{(n)},S_j^m,R_k^m} \prod_{n=1}^{N} (\prod_{m=1}^{M_n} E^m{}_{connection}(S_i^{(n)},S_j^m,R_k^m))^{M_n} \quad (8)$$

where N is the number of stroke groups, M_n is the number of outgoing edges for n^{th} symbol, and $E^m{}_{connection}(S_i^{(n)},S_j^m,R_k^m)$ is edge weight.

Before search is conducted, pruning is done on the weighted direction graph using edge weight in order to improve the search efficiency. Then breadth-first search algorithm with pruning is utilized to find the top N optimal CESGs that represent the structure of the underlying chemical expression.

4. Semantic Verification

The top N optimal CESGs generated by structure analysis are re-ranked using semantic verification. Penalty scores are assigned to those candidate CESGs if they fail the validation.

We conduct grammar checking and valence checking to verify candidate CESGs. For grammar checking, we create context-free grammar based on the syntactical characteristics of chemical expressions. For example, numerical symbols cannot be reactant or product; and chemical elements cannot be subscript or superscript. The context-free parsing algorithm described in [7] is used to analyze CESGs.

5. Experiment

We conducted experiments using the proposed framework on our own database that consists of 35,932 handwritten chemical expressions written by 300 different people. The database contains 2,000 unique chemical expressions copied from the chemistry text books of high school and colleague. The total number of chemical symbols covered in the database is 163 including chemical elements, digits, reaction condition symbols, e.g., '=', '→' and '⟹', operators, e.g., '+', '-', and other frequently used symbols such as '↑', '↓', '%', '℃', etc. Various drawings of bond samples are present in the samples, too. Figure 1 shows some example expressions. About 25% of the samples are organic expressions. We divided the samples into a training set and a testing set with 25,934 and 6,398 expressions, respectively.

In our experiments, we measure recognition accuracy by expression, which is much stricter than measuring by symbol. An expression is considered to be correctly recognized if all of its symbols are correctly segmented and the underlying structure of those symbols is correctly identified. The average length of expressions in our database is about 21 symbols. Each component in our recognition framework was evaluated and the results are summarized in Table 1.

Table 1: Recognition Rate (%)

	Top 1	Top 2	Top 3	Top 4	Top 5
Symbol Grouping	85.9	92.4	94.1	95.3	95.6
Structure Analysis	84.1	85.0	85.5	85.7	85.8
Symbol Grouping + Structure Analysis	74.1	79.8	81.8	82.6	83.0
Symbol Grouping + Structure Analysis + Semantic verification	75.4	80.6	82.2	82.8	83.1

It should be noted that when evaluating the accuracy of structure analysis, correct segmentation results were used in order to decouple the structure analysis component from the symbol grouping component. Experimental results show that semantic verification contributes 1.7% relative accuracy improvement. Overall, our proposed framework achieved Top-5 83.1% recognition accuracy on our large handwritten chemical expression database.

6. Conclusion

We presented a novel unified framework for handwritten chemical expression recognition. A statistical symbol grouping algorithm that handles common symbols and bonds in a consistent manner is proposed. A graph-based representation (CESG) is defined for both inorganic and organic expressions and structure analysis is formulated as a search problem for CESG over a weighted direction graph constructed using statistical modeling that leverages both spatial and contextual information. Semantic verification is used to re-rank top *N* optimal CESGs generated by structure analysis. The experiment results on a large data set show the effectiveness of our framework.

7. Acknowledgement

Thank Yu Zou, Xinjian Chen and Wei Cai for their contribution to this work; and Qi Zhang for proof reading the manuscript.

8. References

[1] Chan K.F., Yeung D.Y. "Mathematical expression recognition: a survey", International Journal on *Document Analysis and Recognition* 3 (2000), pp. 3-15.

[2] Eto Y., Suzuki M., "Mathematical formula recognition using virtual link network", In Proc. of 6th Intl. Conf. on *Document Analysis and Recognition*, 2001, pp.762-767.

[3] Garain U., Chaudhuri B.B., "Recognition of Online Handwritten Mathematical Expressions", IEEE Transactions on *Systems, Man, and Cybernetics*, Part B 34, no. 6, Dec. 2004, pp. 2366-2376.

[4] Education Pack for Windows XP Tablet PC Edition 2005. http://www.microsoft.com/windowsxp/downloads/tabletpc/educationpack/default.mspx.

[5] Microsoft Math 3.0. http://www.microsoft.com/education/products/student/math/default.aspx#overview.

[6] F. Mokhtarian, A. Mackworth, "Scale-based description and recognition of planar curves and two-dimensional objects", IEEE Transactions on *Pattern Analysis and Machine Intelligence*, 1986,Vol. 8, No. 1, pp. 34-43.

[7] J.C. Earley, "An efficient context-free parsing algorithm", Thesis, Carnegie Mellon University, Ph.D. Computer Science Dept., 1968.

[8] N. Hewahi, M.Al Nono, M. Nasar, M. Hamed, H. Hamed, "Chemical Ring Handwritten Recognition Based on Neural Networks", *Ubiquitous Computing And Communication Journal* 3, no. 3, 2008.

[9] Ouyang T.Y., Davis R., "Recognition of Hand Drawn Chemical Diagrams", In Proc. of the National Conf. on *Artificial Intelligence*, 2007, pp.846-851.

[10] K. Toyozumi, N. Yamada, K. Mase, T. Kitasaka, K. Mori, Y. Suenaga, T. Takahashi. "A Study of Symbol Segmentation Method for Handwritten Mathematical Formula Recognition using Mathematical Structure Information", In Proc. of the 17th Intl. Conf. on *Pattern Recognition*, 2004, Vol. 2, pp. 630-633.

[11] Shilman M., Viola P., Chellapilla K. "Recognition and Grouping of Handwritten Text in Diagrams and Equations", In Proc. of the 9th Intl. Workshop on *Frontiers in Handwriting Recognition*, 2004, pp. 569-574.

[12] Tenneson D., Becker S. "ChemPad: Generating 3D Molecules From 2D Sketches", In Proceedings of *SIGGRAPH'05*: ACM SIGGRAPH 2005 Posters. 2005. 8.

[13] J. Ramel, G. Boissier, H. Emptoz, "Automatic Reading of Handwritten Chemical Formulas from a Structural Representation of the Image", In Proc. of 5th Intl. Conf. on *Document Analysis and Recognition, 1999,* pp. 83-86.

[14] J.F. Yang, G.S. Shi, K. Wang, Q. Geng, Q.R. Wang, "A Study of On-line Handwritten Chemical Expressions Recognition", In Proc. of 19th Intl. Conf. on *Patten Recognition*, 2008.