Research on Chemical Expression Images Recognition

Chen Hong, Xiaoping Du, and Lu Zhang

College of Software, Beihang University, Beijing, China 100191

*hcwelldone@126.com

Keywords: image recognition, two-dimensional organic chemical structure

Abstract. There is a large amount of chemical structure stored in image format which is in the scientific literature and the network. Researching on how to recognize chemical structure from images is contribute to storage management of relevant images. The overall recognition effect of chemical structure has room for improvement. This paper proposes a new recognition for two-dimensional organic chemical structure. Firstly, preprocess the image, and then recognize and remove the isolated chemical symbols and dashed bonds. Then vectorize the remaining images and recognize the accretive chemical symbols and chemical bonds from the vectorization results. Finally, repair the results of misidentification and reorganize the chemical symbols and chemical bonds. The experiments show that the recognition rate in this paper is higher than the existing universal recognition method.

Introduction

With the development of computer technology, it is necessary to use the new data format to store the special content of scientific documents, so as to make the contents of the computer readable information, and to facilitate the exchange of information on the Internet. Data format for storing chemical structure information is one of them, such as InChI[1], CML[2]. A large amount of information about chemical structure can be found in the scientific literature and the Internet. These chemical structures are mostly found in the PDF or image files, which are not recognized by the computer and stored in a computer readable format. It takes much time and effort to transform the chemical structure into a computer readable format by artificial means. It is necessary to research the method of automatic recognition and reconstruction of chemical structure in the picture.

At present, the chemical structures of two kinds of sources are identified: one is based on the recognition of strokes [3]. The research in this area is mainly for chemical structure written by handwriting input device. Writing the stroke as the unit to identify chemical symbols and chemical bonds, the chemical structure of the handwritten pictures are automatically converted into chemical files which can be identified by computer. The other is based on the image of the chemical structure of the identification [4]. By processing the chemical structure of the whole image, the chemical symbols and chemical bonds are separated and identified, and the chemical structure information is finally completed. This paper researches on two main problems in the chemical structure image recognition. Firstly, only discuss the chemical bond and the symbol between the intervals of the picture. This kind of image is relatively simple to extract and separate chemical symbols and chemical bonds. For the existence of adhesion, the error separation leads to poor recognition results. Secondly, only deal with the common chemical bond types. It is not complete to identify the real wedge and the virtual wedge. This article proposes a new recognition process to improve the accuracy of the recognition results.

Problem description and solution

Two-dimensional organic chemical structure. Usually, the chemical structure of the research are all two dimensional organic chemical structure as shown in Figure 1. For the convenience of reading, the following are referred to as chemical structure. Usually, the chemical bond is called a chemical structure node, and a chemical structure node can have a number of chemical bonds. Among them, the

node labeled with chemical symbols is called explicit node, and the nodes are not labeled with chemical symbols is called implicit node.

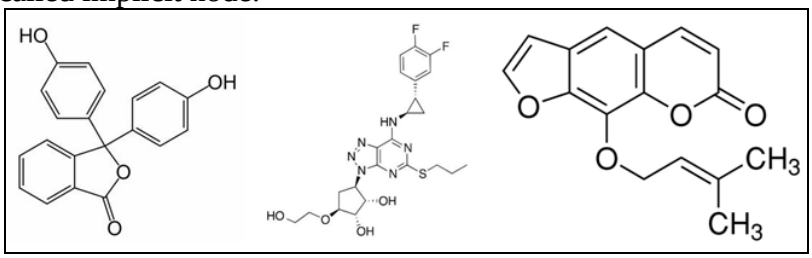

Fig. 1 Chemical structure

Chemical symbols in chemical structures are mainly B, C, F, H, M, N, O, P, R, S, a, e, i, l, r, 2, 3, 4, 5, 6, 7, 8, 9, positive charge (+), negative charge (-).Chemical bond type as shown in figure 2. In figure 2, 1 for single and 2 for double bond, 3 for the triple bond, 4 is a solid wedged bond, 5 for virtual wedge key, 6 and 7 for the benzene ring of two different representations. The form of benzene ring of 7 is composed of three single and three double bonds, can be summed up to 1 and 2.This article only need to deal with the benzene ring's special form of 6.

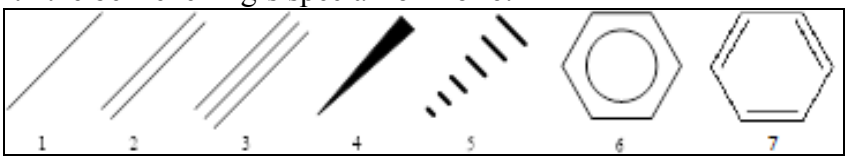

Fig. 2 Chemical bond type

General idea. In this article, the chemical structure information of the image is extracted and displayed in four main stages, which is named as OCSR (Optical Chemical Structure Recognition).

Preprocessing is needed before image recognition. The input images are generally RGB images. This paper is concerned with how to extract structural information from chemical structures. It will simplify the process of recognition to convert the RGB image into two value image. Chemical symbols and chemical bonds are required to be identified after image preprocessing. Chemical symbols are English letters and numbers, which can be identify by OCR technology. Chemical bonds are mainly composed of straight line segments, which can be identify by detecting the linear and the position of the image. Some wrong identified chemical symbols and chemical bonds need to be repaired after identification. Restructuring both, and restoring the original topological relations between them. Finally, the chemical structure of the recombinant is displayed in the chemical structure editor.

Identification of chemical structure

Image preprocessing. In the image preprocessing stage, this article first makes the image a gray-scale map, and then binary the gray-scale map. The gray scale of the image is completed by using the cvCvtColor function provided by the OpenCV library. Binary processing uses Otsu method.

Chemical symbols and chemical bond identification.

Image connected domain separation. Most chemical symbols and chemical bonds are separated from the chemical structure, and only a few are stuck together. Independent chemical symbols can be separated and identified in advance, and these symbols are removed from the original image and then processed. Image connected domain can be found in the following method: In the same domain, the other pixels in the eight neighboring region of any pixel point must be within the same connected domain.

Non-adhesive chemical symbol recognition. The non-adhesive chemical symbol refers to the chemical symbols in chemical structure image which is not connected with the chemical bond.

The high aspect ratios of chemical symbols in chemical structure are in a range. In this article, the high aspect ratio range of chemical symbol connected domain is [0.8, 3.0]. To all the connected domains that satisfy the high aspect ratio, the GOCR[5], OCRAD[6] and TESSERACT[7] are used to identify the open source tool.

Virtual wedge key recognition. Because of the image rendering, clarity and other reasons, the components of the virtual wedge key are not strict in the straight line segment, some of which are the points and the irregular small clumps. In this article, the virtual wedge is identified by the rule of each component of the central point can be connected into a straight line. First, a group of connected domains of virtual wedge keys are divided into one group, and then the linear correlation is calculated by the center point coordinates of all connected domains in the group to determine whether the key is virtual or not.

Image thinning and vectorization result merging. Chemical structure is entirely made up of lines. Over wide lines in the picture does not increase the chemical structure of information, but will increase the difficulty of the combination of vector results. The chemical structure is refined into a single pixel width for subsequent processing. This article adopts Rosenfeld image thinning method. In order to improve the speed of the thinning algorithm, this article adopts the high efficiency algorithm proposed in the literature [8].

Identifying the chemical bonds needs to extract the straight line segments in the image. Converting pixel image into vector map help to identify straight lines and judge their position relationship. In this paper, the image is vectored based on Potrace [9]. Straight line and curve information can be extracted from the result of vectorization.

In the process of vector, Potrace will make the image of the internal and external contour vector. A number of small fold line will produce in the image of the straight line intersection. So the same line segment in the original image can be represented by multiple vector segments. It is needed to extract and merge the chemical bonds and chemical symbols from the line segments. Cut off the line segment from its inflection point, so that the result of the vector is a straight line segment. Set the threshold according to the angle of line segment, and combine the line segments which represent a same straight line into a long straight line segment. This ensures that the same straight line segment in the original image is represented by a vector line segment.

Real wedge key recognition. After image thinning, the real wedge key is a single pixel line segment. In order to separate the real wedge key segment from all straight line segments, the image information is obtained from the original image. After dividing line segments, the linear correlation and variance of line segment width are calculated to judge whether the corresponding straight line segment is real wedge.

Adhesion chemical symbol recognition. In some chemical structures, chemical bonds are linked to chemical symbols, which result in the identification of chemical symbols into multiple chemical bonds.

Adhesion chemical symbols in chemical structure are mainly B, C, F, H, K, N, O, P, S, a, r. After merging the vector results, the original image is converted into a collection of straight lines. The adhesion points in the collection have been separated, but the connecting points of the chemical symbols are also separated. The work to be done is to find a set of chemical symbols from the collection of these line segments.

In chemical symbols, the curve section is a short segment set after the vectorization. Draw these short segments into a graph, separate the connected domain of the graph. Identify connected domains which meet the conditions described in the before section, and the identification of the symbols is kept as chemical symbols. For connected domain which is not recognized, combine it with a long straight line segment which is connected with it, and then identify again.

Chemical symbols and chemical bond repair and recombination.

Repair double characters chemical symbol. In the before section, the “i”, “l”, negative charge (-) and the chemical bond cannot be distinguished between the straight line segment, to prevent the chemical bonds to identify these chemical symbols, these chemical symbols are placed in the list of the identification results.

In order to ensure the correctness of the results, the results of these are ignored. The elements of “i” are Bi and Si, and the element of the “l” is Cl, and the negative charge (-) exists in the left or right side of the chemical symbol. This paper calculated the single slope to judge and repair double characters of chemical symbols with threshold.

“Super” atomic merging. For multiple horizontal or vertical chemical symbols, they need to be merged into a "super" atom. In chemical structure, a "super" atom is a whole being connected with other chemical symbol nodes. If the horizontal or vertical chemical symbols are merged into a "super" atom, they are treated as independent chemical symbols, and the chemical symbols and chemical bonds are combined. In this paper, set the threshold by calculating the distance between the external rectangle of the adjacent connected domain, make a combination of connected domains which the distance between them is less than the threshold.

Removal of non-chemical bonds. If the image aliasing is serious, it will produce more small line segments in the vector. In order to reduce the recognition error, the chemical bonds in the chemical bonds can be identified.

In this paper, the chemical bond and the non-chemical bonds are distinguished according to the length threshold. In the use of the length threshold, the short chemical bonds cannot be recognized as a non-chemical bond to remove, because the chemical structure of the image can also have a very short chemical bond. By observing the chemical structure of the image, a fact that short chemical bonds are always connected to the two chemical symbols is found. Further distinguishing work can be done According to this point.

Structured output and display. In this paper, a JavaScript based open source chemical structure editing tool JSME[10] is used to demonstrate the results. JSME can read a string according to the string to display the corresponding chemical structure.

In this paper, the adjacency matrix of a chemical structure identification result is converted into a character string, and then is sent to JSME to display.

Analysis and discussion

Currently there are two available chemical structural image recognition tools. One is the only one of the commercial chemical structure image recognition tool CLiDE, the other is the only one free online chemical structure image recognition tool OSRA. In the literature [11], the recognition effect of CLiDE and OSRA was compared. The results showed that the recognition effect of OSRA was better than CLiDE. In this article, the recognition results of OCSR and OSRA are compared.

In this article, 200 chemical structures are found, which are identified by OCSR and OSRA, and then are measured the similarity between them by identifying the results of the OSRA and OCSR using Tanimoto Coefficient [12]. Tanimoto Coefficient is originally used to measure the similarity of two sets A, B. In this article, the chemical structure is converted into a set of elements, and the elements of the set are of all chemical bonds and all chemical symbols.

Table 1 shows the recognition rate between OSRA and OSCR. From the table, we can see that the ratio of OCSR is not very high, but most of the chemical structure images can be recognized of the Tanimoto Coefficient higher than 85%, which can help simplify the work of chemical structure into a computer.

Table 1 Comparison of OSRA and OCSR (200 chemical structures)

	MVC	T=100.0%	T>95.0%	T>90.0%	T>85.0%
OSRA	82.2%	71	82	105	114
OCSR	90.0%	95	104	139	157

Summaries

In this paper, the research status of chemical structure image recognition is investigated, and the current problems in chemical structure identification and image based chemical structure identification are summarized. In order to solve the problem of recognition of the chemical structure of the adhesion and the recognition effect of the real wedge key and virtual wedge key, a new recognition process is proposed to improve the recognition rate of chemical structure.

In this paper, the recognition effect of OCSR is compared witha free online chemical structure image recognition tool OSRA. 200 chemical structures are selected from the network, which are

identified by OCSR and OSRA, respectively. The results show that the overall recognition effect is better than OSRA.

References

[1] Heller S R, Stein S E, Tchekhovskoi D V. InChI: Open access/open source and the IUPAC international chemical identifier[C]//ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SO-CIETY. 1155 16TH ST, NW, WASHINGTON, DC 20036 USA: AMER CHEMICAL SOC, 2005, 230: 1025.

[2] Murray-Rust P, Rzepa H S. Chemical markup, XML, and the Worldwide Web. 1. Basic principles[J]. Journal of Chemical Information and Computer Sciences, 1999, 39(6): 928.

[3] P. Tang, S. C. Hui, C. W. Fu. Online chemical symbol recognition for handwritten chemical expression recognition[C]//Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on. IEEE, 2013: 535.

[4] M.E. Algorri, M. Zimmermann, M. Hofmann-Apitius. Automatic recognition of chemical images[C]//Current Trends in Computer Science, 2007. ENC 2007. Eighth Mexican International Conference on. IEEE, 2007: 41.

[5] Optical Character Recognition (GOCR). http://jocr.sourceforge.net/download.html.(accessed February 20, 2014).

[6] Ocrad-GNU Project-Free Software Foundation (FSF).http://mirror.bjtu.edu.cn/gnu/ocrad/.(accessed February 20, 2014).

[7]Tesseract-OCR.http://code.google.com/p/tesseract-ocr/downloads/detail?name=tesseract-3.02.02-win32-lib-include-dirs.zip.(accessed February 20, 2014).

[8] Cychosz, J. M. Efficient binary image thinning using neighborhood maps. In Graphics gems IV; Academic Press Professional, Inc.: San Diego, CA, 1994; pp 465-473.