Chemical Structure Recognition (CSR) System: Automatic Analysis of 2D Chemical Structures in Document Images

Syed Saqib Bukhari*, Zaryab Iftikhar†, Andreas Dengel*†
*German Research Center for Artificial Intelligence (DFKI), Germany
†University of Kaiserslautern, Germany
saqib.bukhari@dfki.de, iftekhar@rhrk.uni-kl.de, andreas.dengel@dfki.de

***Abstract*—In this era of advanced technology and automation, information extraction has become a very common practice for the analysis of data. A technique known as Optical Character Recognition (OCR) is used for recognition of text. The purpose is to extract textual data for automatic information analysis or natural language processing of document images. However, in the field of cheminformatics where it is required to recognize 2D molecular structures as they are published in research journals or patent documents, OCR is not adequate for processing, as chemical compounds can be represented both in textual as well as in graphical format. The digital representation of an image based chemical structure allows not only patent analysis teams to provide customize insights but also cheminformatic research groups to enhance their molecular structure databases, which further can be used for querying structure as well as sub-structural patterns. Some tools have been made for extraction and processing of image-based molecular structures. Optical Structure Recognition Application (OSRA)[11] being one of the tools that partially fulfill the task of recognizing chemical structural in document images into chemical formats (SMILES, SDF, or MOL). However, it has few problems such as poor character recognition, false structure extraction, and slow processing. In this paper, we have developed a prototype Chemical Structure Recognition (CSR) system using modern and advanced image processing open-source libraries, which allows us to extract structural information of a chemical structure embedded in the form of a digital raster image. The CSR system is capable of processing chemical information contained in chemical structure image and generates the SMILES or MOL representation. For performance evaluation, we have used two different data sets to measure the potential of the CSR system. It yields better results than OSRA that depict accurate recognition, fast extraction, and correctness of great significance.**

***Keywords*-Document Analysis Systems, Chemical Structure Recognition, Information Extraction, Cheminformatics, Symbols and Graphics Recognition**

I. Introduction

The abrupt rise of modern computer technologies has highlighted the need for new machine parsable formats for processing information, within the context of document image analysis. Formats such as SMILES (Simple Molecular In Line Entry System)[6], SDF (Structure Data File), MOL (Molecular File)[9], CML (Chemical Markup Language)[9] etc. have appeared, which are appropriate for representing molecular structure information. However, the excess of literature in the field of cheminformatics that existed before the development of these formats does not acquire such established machine-parsable formats for representing molecular information. The necessity for automated extraction and parsing of chemical structures has proved to be very challenging, such that even after the development of several tools for chemical structure recognition, none has achieved global acceptance.

In scientific literature, tremendous amount of information is related to cheminformatics that needs to be compiled and later used for querying. Many new tools have been developed with the requirement to parse chemical structures in the form of raster image representation. The basic purpose is to parse this chemical information and make it globally available for searching for its relevance in any specific discipline. Once a reliable application is built for digitizing image based chemical structures, it can then be used to process chemical literature and store the outcomes in a universal database, as illustrated in Figure 1.

Two examples of globally available databases that links chemical structures with bio-medical targets are PubMed[13] and PubChem[14]. The later has over 19 million chemical structures with a cross-reference link to bio-assay data, similar structures, and medical research descriptions. If a universal database can be made from these information resources, aggregating all identified chemical structures with their linkages to biochemical pathways, disease states, and therapeutic applications. With this information an advanced tool can be established for both drug discovery and bio-medical research as stated by Jungkap Park et al.[16].

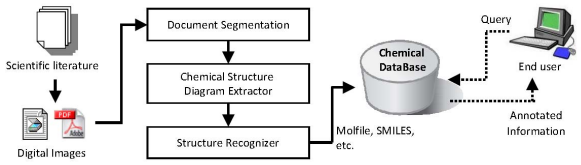


Figure 1: Structure Recognition Overview Diagram (Source: Shedden et al.[16]).

Nowadays there is a necessity for automated processing of chemical structure images in research articles. However, the images in scientific journals are encoded in various

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00230

digital formats and these images might also contain embedded text segments that does not represent anything related to chemical structure and needs to be filtered off. There is a need for a mechanism to extract text as well as shapes/polygons that made up to form 2D chemical structure. In this paper, we are proposing the Chemical Structure Recognition (CSR) system. It allows us to extract molecular structures as they appear in patent documents, journals, and images. Once the structure is extracted, it performs a series of operations that fetches chemical information contained in the form of bonds, atoms, rings and charges etc. Finally this connectivity information about the molecule is represented in SMILES format.

The structure of this paper is as follow. Section II describes the literature review in terms of existing open-source tools (such as OSRA[11]) and commercial tools (such as CLiDe[12]) for recognizing chemical structures in document images, and the standard formats for chemical structure representation. It also briefly states the areas in which enhancement can be made. Section III described the technical details of the proposed Chemical Structure Recognition (CSR) system. In Section IV we present the performance evaluation of the CSR system an its comparison with respect to the open-source OSRA tool. Finally, Section V summarizes the conclusion and presents some future enhancement ideas.

II. Literature Review

Here we will consider few state-of-the-art chemical structure recognition tools. This will help in the better realization of the motivation behind the development of the proposed CSR system. We will also cover few of the Standards such as SMILES and MOL used for structure representation.

OSRA[11]: It is an open-source chemical structure recognition tool developed by Igor Felippov[11]. It is a utility designed to translate raster image based representation of chemical structures, as they appear in scientific journals, patent documents etc. into a computer chemical structure format. OSRA uses GraphicsMagick[5] that allows it to read a document in almost all major image formats, then generate the SMILES notations or MOLfile of the molecular structure images encountered within that document. The development of OSRA started way back in 2007 and it has been funded with federal funds from the US National Cancer Institute, National Institutes of Health under contract N01-CO-12400[11].

OSRA is a command line tool along with an average web interface. It is written in C++ and uses CImg[1] library for image processing and GOCR[4] for text recognition. The later has not seen any update in past five years. Which means, in character recognition methodologies of the process, improvements can be made using a recent OCR library with active development such as TesseractOCR[3]. The computation of nxn distance matrix of connected components, which is used for page segmentation in OSRA, makes the entire processing very slow if a page has huge paragraphs of text as compared to a page having only structures.

CLiDE[12]: It is a chemical OCR tool aimed at automated extraction of chemical information from either electronic PDF version or from the printed chemistry literature. CLiDE development started back in 2008 and it has been commercially available in three versions: 1) CLiDE Standard (one structure recognition per process), 2) CLiDE Professional (a large number of chemical structures per process), and 3) CLiDE Batch (supports batch processing and web service environments).

CLiDE, based on its version, supports command line tool as well as a GUI for Windows operating system. It is written in C++ and the entire procedure seems fluid. However, CLiDE needs document images of at least 300 dpi resolution to achieve reliable result. However, the embedded molecular diagrams in document usually have 72–96 dpi. Due to license requirements caveat and the tool being not open-source, we could not sort out which Image processing and OCR engines were used in CLiDE. Moreover, we were also not able to test the data sets on this tool, therefore, the evaluation Section IV only includes comparison with OSRA.

Standards for Representation (SMILES and MOL): SMILES and MOL are the chemical formats that contains information about a chemical structure that is machine readable. This processed data can be stored in databases for global usage across various disciplines.

SMILES[6] is short of Simplified Molecular Input Line Entry System, which is a linear string notation for representing molecules and reactions. SMILES comes with a simple vocabulary and only a few specification rules. SMILES notations referring to a chemical structure can, in turn, be utilized as "word" in the vocabulary designed for storing information about chemicals. SMILES notation is a concatenation of a series of characters containing without spaces. Hydrogen atoms may be omitted or included depending on the configuration. By applying the basic specification rules, SMILES can be generated/read as illustrated in Figure 2.

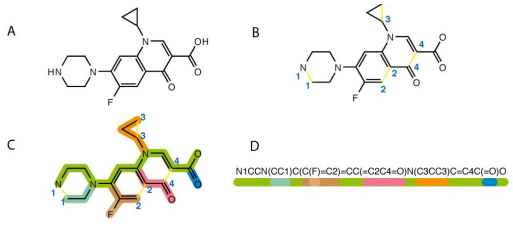


Figure 2: Rules for reading a SMILES [2].

MOL[9] is a chemical file format for storing connectivity patterns of atoms and bonds in a molecule. This data structure contains some header information in the form

of a connection table (CT). The CT comprises of atom information, coordinates, bond types and involved atoms indices, stereo chemistry followed by sections for complex information. MOLfile format is precisely the most common, and almost all cheminformatics tools are able to read the format. Figure 3 describes the format / structure of a MOLfile for benzene.

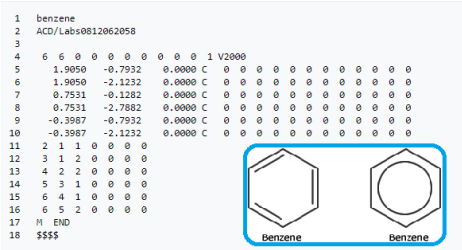


Figure 3: Structure of MOLfile[18].

III. The Proposed Chemical Structure Recognition (CSR) System

We divide the entire process for translating a digital image based chemical structure to machine-readable format into two sub-processes. Extraction process, which is responsible for extracting chemical structures images from a document image. The second is the Recognition process, which takes each individual chemical structure image, applies a series of computer vision techniques to achieve the desired chemical format. The entire development of the CSR system is based on open-source libraries as follow: i) OpenCV 3.3 with Python binding[8], ii) Open Babel 2.4[15], iii) Tesseract OCR 3.05[3], and iv) Python Libraries (Numpy +mkl, Scipy & Scikit-Image, Matplotlib, Pytesseract, PIL, Flask and its dependencies). The following sections describe in detail each step for both of the extraction and recognition processes.

A. The Extraction Process

Here our task is to segment the text contents and graphics contents (containing chemical structures) from a document image. Thus making it possible to extract all individual chemical structure images. The extraction process consists of the following stages to achieve its results.

Image Pre-processing: This stage includes basic computer vision techniques to prepare the document for further processing. OpenCV provides built in one liner functions to achieve gray-scaling, blurring, and thresholding.

Run Length Smearing: Once the document is pre-processed we execute smearing algorithm in parallel execution. This algorithm creates two separate smeared images, horizontally smeared and vertically smeared. The two images are then logically ORed to complete the smearing process. This process merges all pixels that belong to a chemical structure into a single connected component.

Labeling and Text-Region Filter: The resultant document from the output of smearing process is further processed to find labels using connected component approach. The text-region filter takes each connected component region, and looks up the same region within the original document and extract all images. Once it has all view-ports it re-performs the labelling on each image to filter out the invalid images those have all small components i.e. text only components.

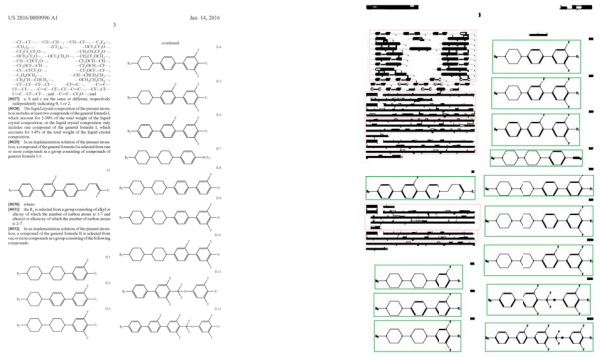


Figure 4: Left: A sample document image containing text and graphics (chemical structures) regions. Right: After the extraction process (that is described in Section III-A) where valid graphics regions are shown with green bounding boxes and invalid text regions are shown with red bounding boxes.

A sample image and its corresponding result after the extraction process is shown in Figure 4, where valid i.e. graphics/chemical-structure regions are shown with green color and invalid i.e text only regions are shown with red color.

B. The Recognition Process

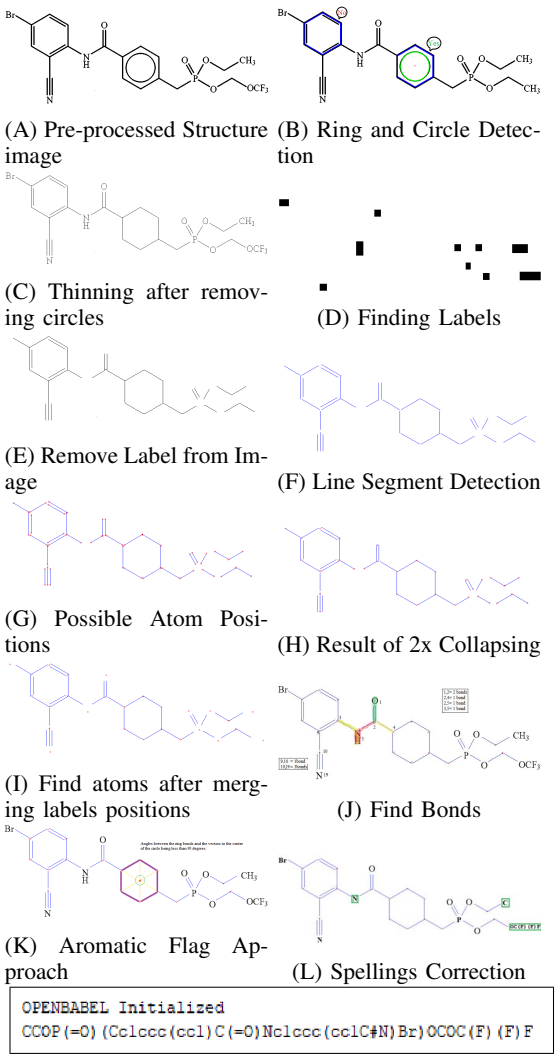
The recognition process works on individual image containing a chemical structure coming from the outcome of the extraction process. In this process we perform a series of operations to decomposes the image into various graphics components for discrete identification such as Double bond, Atom labels, Old aromatic bond, Triple bond, Super-atom labels, Single bond and Carbon Atoms if no label.

A complete step by step diagram of recognition process in shown in Figure 5. The Recognition process consists of the following stages to achieve its results.

Ring Detection: The first step of the decomposition stage is to detect hexagonal rings and triangular wedges, using the contour logic provided by OpenCV. Once the rings are detected we performed Hough Circle[7] based on the dimension of the rings region and store the center and radial information about the circles that represents aromatic bonds (see Figure 5B). At the end of this process we removed the circle and return an image without circles.

Thinning: Once the circles are removed we then performed morphological thinning[19] on the image (see Figure 5C).

Finding Labels: In this stage we find connected components of the size of letters. We merge the nearby components (see Figure 5D) and then perform OCR operation to detect



(A) Pre-processed Structure image

(B) Ring and Circle Detection

(C) Thinning after removing circles

(D) Finding Labels

(E) Remove Label from Image

(F) Line Segment Detection

(G) Possible Atom Positions

(H) Result of 2x Collapsing

(I) Find atoms after merging labels positions

(J) Find Bonds

(K) Aromatic Flag Approach

(L) Spellings Correction

```
OPENBABEL Initialized
CCOP(=O)(Cclccc(ccl)C(=O)Nclccc(cclC#N)Br)OCOC(F)(F)F
```

(M) Recognition Process Output

Figure 5: The Recognition Process Steps (described in Section III-B).

the characters of each connected component. The OCR operation uses a white list configuration for acceptable characters. We also filter off few characters like “, . :” that can be stated as mistakes. This operation returns a list of labels with their box coordinates and the label itself.

Remove Labels from Thin Image: We perform a series of regular expression on the resultant label list to find labels that are only alphanumeric or have alphanumeric with $+, -, (,)$ symbols. The following symbols are used to represent a positive or negative charge on atoms. The rest of the elements are removed from the list. Using this approach we are only left with coordinates that allude to possible atom labels or super-atom labels. Once done, we remove these regions from the thin image as shown in Figure 5E.

LSD on Thin Image: For finding bond information, we execute the Line Segment Detector (LSD) [17] algorithm on the thin image in order to detect lines (see Figure 5F). As the circles and characters are already removed so this process is straightforward. The output of LSD algorithm goes through the post-processing stages for solving possible errors and for acquiring accurate bond detection as follows: i) keep only perfect horizontal and vertical lines, ii) remove double lines, and iii) correct/adjust those lines which are closely parallel to each other.

Find Atoms: Using the lines and label information we find possible positions of atoms within the structure. There are two possible positions for atoms: i) At the edge of each line, and ii) Label regions. As we have the list of lines and label regions, we do not directly create atoms, but we first find possible atom position. This process is straight forward as it performs the following steps to make a list of atom positions: i) Marks the edge position of each line (see Figure 5G), and ii) Collapse the marked positions twice (see Figure 5H). In the next step, we take the atom positions as well as central positions of the labels and merge them. So that the alleged atom position with nearby labels become one as illustrated in Figure 5I. Once the positions are finalized we create an atom data structure for each position.

Find Bonds: Once all the atoms are created we need to create bonds. The list of lines returned by the LSD algorithm can be used to detect bonds but it is not as easy as it looks like. We need to have the right combination of atoms and the number of bonds between them. We used a data structure fulfill this requirement. We iteratively go through the list of lines and for each endpoint of the each line we find the nearest atom by iteratively going through the list of atoms. We compare the positions of atoms with the endpoints of the line. At the end, we get a new list that marks the number of lines between each atom pair, which actually represent the number of bonds.

Open Babel Pre-processing: After finding all bonds and atoms we need to mark few flags. The Open Babel[15] requires the meta-data about atoms and bonds for proper translation. Therefore we need to further process the information to find aromatic(see Figure 5K), terminal, and up-down bonds. We also need to make label corrections (see Figure 5L) and find charges on atoms. Hence the decomposition stage ends.

Assemble Structure: This is the final step of the recognition process. The decomposition stage harvests two data structures, one for atoms and the other for bonds. They are used to map the connectivity information to underlying the Open Babel API. Based on the provided data an Open Babel molecule object (OBMol) is created to assemble the molecular structure. For each atom, a look-up operation is performed in the Open Babel’s periodic table to fetch its atomic number. If an atom has no element then carbon atom is assumed in this case and its atomic number is inserted. At the end, for every single atom, an OBAtom object is created. Once all atoms are defined in the OBMol object.

All bond information from the bonds data structure needs to be added. This meta-data of bonds sets the flags that help the Open Babel to generate a molecule file that depicts the exact orientation of the structure with respect to original structure image. Bonds are created using OBBond object and are thus aggregated into the main OBMol object. Super-atom itself can be considered as a molecule, we need to decompose it into atoms and bonds objects and then store them one by one in the main object. Then Open Babel processes the object and generates a SMILES string. Figure 5M shows the SMILES notation which is generated as the outcome of the recognition process for Figure 5A.

IV. Performance Evaluation

In order to test the performance of the presented CSR system with respect to OSRA[11], we used a data set of 50 single structure images and 7 multi-structure document images with 89 embedded chemical structures. We performed the analysis between the two tools on the same system so that the outcome seems reliable. Specifications of the system that was used for testing are Intel Core i-3 2nd Generation 64-bit, Ubuntu 16.04 with 8 GB RAM. The only way to make this tool more competitive and globally acceptable is to check various aspects of performances. The necessity of these analysis yields results about the efficiency of this tool. Following are the type of analysis that was conducted to verify the performance of the CSR system: i) Batch Execution Time, ii) Batch Structure Extraction Accuracy, iii) Multi Structure Recognition Accuracy, and iv) Single Structure Recognition Accuracy.

Batch Execution Time: Since OSRA and CSR both support batch processing, which is processing of multi-structure images or multi-page PDF documents. We analyzed the execution time of the batch process on 7 document images, on the whole, the process examined round about 90 embedded chemical structure in the raster images. See Figure 6 for the result of the experiment. The results clearly yield that the CSR system outperforms OSRA in the department of speed. We can infer that its execution time is 36% faster than OSRA on the given data set. The reason for slow processing of OSRA is its nxn distance calculation for extraction of structure, whereas the CSR system uses a faster run length smearing approach.

Multiple Structure Extraction Accuracy: Next, we compared our CSR application with OSRA for structure extraction. This analysis is about structure extraction from multi-structure images. The purpose is to check whether the tools are capable of extracting all the chemical structures or they left out few of the structures in the data set. Figure 7 yields the graph for this analysis. The line on the chart represents the actual number of chemical structures in each of the document. We can clearly see that the CSR system is capable of extracting all structures in all documents with

Figure 6: Execution time comparison between the presented CSR system and the state-of-the-art OSRA tool.

an astounding 100% accuracy. Whereas OSRA detected one less in $doc5$ and one more in $doc4$.

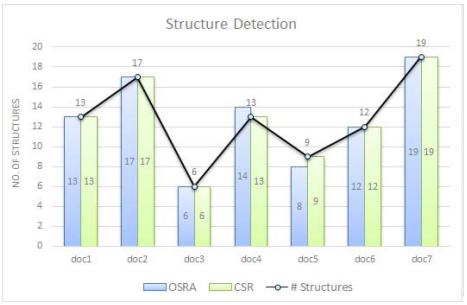


Figure 7: Structure extraction comparison between the presented CSR system and the state-of-the-art OSRA tool.

Multiple Structure Recognition Accuracy: The essential requirement of a chemical recognition tool is to recognize the structure correctly without manual intervention. Because it is the correct translation that the research groups of cheminformatic rely on not the speed. Figure 8 portrays the analysis for correctness. The line on the graph represents the actual number of structures, whereas the bars represent the number of accurately recognized structures by the respective tool. In this experiment, the correctness criteria were defined in terms of perfect SMILES generation from structural information. The structure was considered to be incorrect if there exists even a single OCR or a single bond error. In $doc3$ and $doc4$ the CSR system takes a huge lead with 5/6 and 11/13 correct recognition. On the whole, it is visible that on average the overall correctness of the CSR system is way ahead of OSRA for the given data set.

Single Structure Recognition Accuracy: In this experiment, we took 50 individual images of different chemical structures. We performed a 1-1 analysis of OSRA's and our methods on the data set. The correctness criteria was the same, as defined in previous section. Success-rate S_{rate} is defined as the percentage ratio of all incorrect recognition I_{recog} to total structures T_{struc} as shown in Eq 1.

$$S_{rate} = \frac{I_{recog}}{T_{struc}} * 100 \tag{1}$$

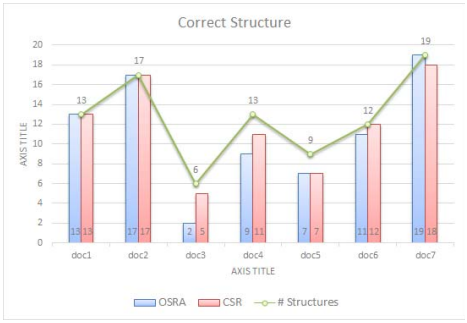


Figure 8: Structure recognition comparison between the presented CSR system and the state-of-the-art OSRA tool.

Table I: Single Structure Analysis on the Accuracy of CSR with respect to OSRA.

Tools	Total Samples	Total Incorrect	Success-rate
CSR	50	7	86%
OSRA	50	13	74%

Table I lists the experiments results. We can clearly see that CSR's recognition-rates are higher than OSRA for the given data set. Most of the images were specifically converted using GraphicsMagick[5] for OSRA because it was not recognizing. Even after conversion of the entire data set, there were still cases, when OSRA was not able to recognize anything and gives "no structure found" error. This clearly shows that even a mature structure recognition tool like OSRA can behave abnormally and this is the reason why structure recognition has not yet received any standard solution.

V. Conclusion

The presented work is the first implementation of the CSR system which is to be considered as a prototype. This system was made entirely from scratch and has the potential to be improved with future upgrades. The main purpose was to come up with an approach that yields not only better but efficient structure recognition results. So that, in near future this system can be further optimize and is welcomed by cheminformatics research groups. The performance evaluation results for execution time and accuracy also show that the CSR system is better than the state-of-the-art tool OSRA [11]. Additionally, on the basis of performance evaluation results, it can also be concluded that the use of run length smearing[10] along with a text-region filter speeds up the extraction process of chemical structures, compared to OSRA's nxn distance matrix of connected components. The use of TesseractOCR[3] instead of GOCR[4] improves the character recognition by a significant fraction. Line segment detector (LSD) oppose to vector tracing which is used by OSRA, is straightforward and does not need much post-processing. Development of tools like OSRA[11] and CLiDE[12] started way back in 2007 and 2008 respectively, with decades of effort and research those tools have been matured. However, development of the CSR system took roughly six months, and it is still capable of defeating OSRA for the provided data set. This clearly portrays this system can mark its name in the list of structure recognition tools in near future.

References

[1] "Cimg toolkit," http://cimg.eu/.

[2] "SMILES," *Wikipedia*, "https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system".

[3] "Tesseract OCR," https://github.com/tesseract-ocr/tesseract.

[4] "Optical Character Recognition (GOCR)," 2000, http://sourceforge.net/projects/jocr/.

[5] "GraphcisMagick image processing library," 2002, http://www.graphicsmagick.org/.

[6] "Open SMILES specification," 2007, http://opensmiles.org/opensmiles.html.

[7] "Hough Circle Transform," *OpenCV 3.4 documentation*, 2017, https://docs.opencv.org/3.4.0/d4/d70/tutorial_hough_circle.html.

[8] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 2000.

[9] A. Dalby *et al.*, "Description of Several Chemical Structure File Formats used by Computer Programs Developed at Molecular Design Limited," *Journal of Chemical Information and Modeling. 32(3)*, 1992.

[10] S. Ferilli *et al.*, "A Run Length Smoothing-Based Algorithm For Non-Manhattan Document Segmentation"," 2012.

[11] I. V. Filippov and M. C. Nicklaus, "Optical Structure Recognition Software To Recover Chemical Information: OSRA — An Open Source Solution," *J Chem Inf Model 49(3):740–743*, March, 2009, https://cactus.nci.nih.gov/osra/.

[12] P. Ibison *et al.*, "Chemical Literature Data Extraction:The CLiDE Project," *J Chem Inf Comput Sci 1993, 33:338-334*.

[13] NCBI, "Pubmed," 1996, https://www.ncbi.nlm.nih.gov/pubmed/.

[14] ——, "Pubchem," 2004, https://pubchemdocs.ncbi.nlm.nih.gov/about.

[15] N. M. O'Boyle *et al.*, "Open Babel: An Open Chemical Toolbox," *Journal of Cheminformatics 3:33*, 2011.

[16] J. Park *et al.*, "Automated Extraction of Chemical Structure Information from Digital Raster Images," *Chemistry Central Journal*, 2009.

[17] R. G. von Gioi et al, "LSD: a Line Segment Detector," *Image Processing On Line, pp. 35–55. DOI: 10.5201/ipol.2012.gjmr-lsd*, 2012.

[18] Wikipedia, "Chemical Table File," https://en.wikipedia.org/wiki/Chemical_table_file.

[19] T. Y. Zhang and C. Y. Suen, "A Fast Parallel Algorithm for Thinning Digital Patterns," *Communications of the ACM, Volume 27, Number 3*, 1984.