OPTICAL CHEMICAL STRUCTURE RECOGNITION

A report on
Computer Vision Lab Project
[CSE-3181]

Submitted By
NEHAL CHANDAN MURDESHWAR - 210962021
JAYASURYAN MUTYALA – 210962009
ABHIRAM REDDY KONDA - 210962003

INSPIRED BY LIFE
MANIPAL
ACADEMY *of* HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY,
MANIPAL ACADEMY OF HIGHER EDUCATION
NOVEMBER, 2023

Optical Chemical Structure Recognition

Nehal Chandan Murdeshwar[1], Abhiram Reddy Konda[2], Jayasuryan Mutyala[3]

[1]MIT Manipal, India

[2] MIT Manipal, India

[3] MIT Manipal, India

[1]nehalmurdeshwar@gmail.com; [2] abhiramrk123@gmail.com; [3] jsmu.dev@gmail.com

Abstract— Optical chemical structure recognition is the process of converting a graphical representation of a chemical molecule into its traditional structural representation. In particular, the network structure should be recognised by the chemical structure identification method with precise atomic/group labels for each node and bond type for each vertex. Considering the additional difficulties caused by hand-drawn molecules, we must factor in the added difficulties and accept that the accuracy may be hindered. Using basic text recognition and corner detection techniques, we first identify the atoms and groups that comprise the nodes of the chemical structure graph. We discover that the performance of the corner detection technique used outperforms that of the line vectorization methods frequently used in other systems. Using a Hough transform, the presence of bonds between the nodes is identified. This approach differs primarily in that we apply a novel paradigm based on several feature descriptors of sliding-window cross-sections. In addition to the conventional method of classifying bonds using the Hough transform, we also use local maxima detectors on single-pixel slices of bond cross-sections and histogram of oriented gradients (HOG) features of wider bond cross-sections. Upon assessing our code with a manually created dataset of 360 basic compounds and contrasting the results of these feature descriptors, we discover that this bond identification method gives us good enough results to be optimistic about future work being done on it.

I. Introduction

In many disciplines, including biology, chemistry, and medicine, the structural diagram—which includes all the chemical details about a particular molecule but is unsuitable for computer analysis—remains the standard representation of organic chemistry data. The task of optical structure recognition, which converts these structures' images into machine-readable, annotated graphic formats that are useful, is still extremely difficult and frequently imprecise. Scientific research would significantly improve not just in chemistry, chemical biology, medicine, and many other subjects, but also in the broad availability of such scientific patents, journal articles, textbooks, and other printed materials. A chemical structure identification tool would also create new prospects for data mining and artificial intelligence in the present day by applying it to datasets that currently only exist as photos. Moreover, no practical effort has been done to extend optical structure detection to chemical compounds that are hand-drawn. The current methods of study concentrate on increasing the accuracy of computer-generated algorithms for very large molecules, which are usually described in scientific patents. Not much work has been done on the smaller, more manageable hand-drawn molecules. Smaller hand-drawn molecules provide distinct and different challenges for computer vision than computer-generated molecules, which are notoriously hard to recognise.

II. LITERATURE REVIEW

The field of optical chemical structure recognition (OCSR) represents an important intersection between chemistry and computer science, in which the conversion of graphical representations of chemical structures into formats that can be Machine understanding is the main goal. This literature review evaluates the methods and technological frameworks of two pioneering OCSR tools, OSRA and CLiDE, which have contributed significantly to the discipline.

CliDE [5], standing for Chemical Document Data Extraction, is a notable commercial OCSR tool that has been evolving since its creation in 2008. It is available in three versions: Standard, Professional, and Batch, each tailored for the Windows operating system with command line or GUI adaptability. However, the effectiveness of CLiDE hinges on high-resolution input, often surpassing the standard resolutions found in scientific papers, exposing a limitation in the tool's adaptability. Despite its proprietary nature—which restricts a full evaluation of its technological framework—CLiDE provides a benchmark for OCR capabilities when compared to its open-source counterparts. Within the OCSR tools, the SMILES and MOL formats are crucial, as they allow for the standardized, machine-readable presentation of complex chemical information, which is essential for interoperability in scientific applications.

OSRA[7], introduced in 2009, is an open-source tool that is distinguished by its professional binarization and image segmentation techniques, which are particularly adept at recognizing aromatic and polybonds. In parallel, ChemReader[6] emerged with a strong preprocessing suite that included noise filtering and size normalization, utilizing a modified Hough transform for line detection which is central to determining bond order and stereochemistry. MolRec, also launched in 2009, originated from the University of Birmingham and showcases a complex rule-based system adept at recognizing and annotating chemical diagrams. Both OSRA and ChemReader have made significant strides in their respective areas, though access to ChemReader and MolRec is limited, which constrains their broader evaluation and impact.

The introduction of Imago in 2011 added to the suite of OCSR tools with its rule-based approach for identifying connections after blurring and binarizing images. Imago's[13] performance is bolstered by a comprehensive dictionary of metaatoms and common chemical abbreviations, showcasing its versatility. Meanwhile, eChem[13] serves a unique educational purpose, providing a user-friendly platform for analyzing chemical structures. It excels at converting raster images into machine-readable formats and relies on the Microsoft Office Document Image Library for the accurate recognition of characters and links, translating them into strings for educational analysis.

DECIMER[14] stands out for its innovation in chemical structure segmentation, utilizing the Mask R-CNN architecture combined with TensorFlow, and is a testament to the spirit of open-source development with resources available on platforms like GitHub. The DECIMER Image Transformer leverages a curated dataset from PubChem, which includes a vast collection of molecules filtered by molecular weight and stereochemical integrity. This dataset is foundational for developing highly accurate OCSR systems and is a clear illustration of DECIMER's commitment to accuracy and data diversity.

RanDepict[13] plays a crucial role in data diversity, adjusting representation parameters to produce spectra representing chemical structures. This tool integrates with established chemistry software suites like CDK, RDKit, and Indigo, enriching OCSR research and development. The integration of tools like RanDepict and DECIMER with platforms such as Google Cloud Platform showcases a blend of scientific precision and creative expression, as seen in the visual arts and animation, reflecting a broader narrative of technological integration and collective advancement across disciplines.

The ORCS[16] report marks a significant advancement in OCSR, presenting a sophisticated method for interpreting molecular structures from images. It introduces a nuanced grayscale conversion technique, a new "noise factor" calculation for noise reduction, and leverages the Potrace library for vectorization, crucial for accurate bond and atomic position detection. The creation of the "CEDe" dataset, with over 700,000 annotated chemical entities, is a milestone for OCSR, enhancing the accuracy of chemical structure reconstructions. The effectiveness of these methods is validated by their success in handling line-touching characters and overall recognition accuracy, positioning the ORCS report as a pivotal contribution to the field.

III. METHODOLOGY

A. Summary

Our methodology for identifying the chemical structures within molecules follows a structured sequence, which is detailed below:

- Utilization of scale-invariant template matching for text label recognition
- Text removal from the image
- Detection of corners
- Detection of bonds
- Classification of bonds.
- Associating corners with atoms and molecular groups

Given the restricted number of labels and the lack of hand-drawn chemical structure data, we used a relatively simple scale-invariant template matching method to reliably detect text in photos. Alternative approaches were investigated, including classifiers that use the histogram of oriented gradients (HOG) and Google Tesseract. Tesseract's broad-spectrum use parameters, designed for structured text recognition, made its assumptions and language models unsuitable for our specific requirements, making its configuration difficult. Although an advanced OCR engine might be considered for future enhancement, it proved impractical at this stage. Similarly, HOG-based classifiers encountered numerous false positives, attributed to the dearth of negative examples in our training set. This model has potential, provided a more substantial training dataset becomes available.

For the identification of key structural features, we applied bond and corner detection to pinpoint areas of interest, such as line intersections indicative of carbon atoms or line terminations at text boxes representing different atoms or functional groups. Prior to line detection via a Hough transform, a Gaussian filter was employed to smoothen the image.

Subsequently, these identified points were used to map out the molecules' atoms, groups, and bonds. The last step involved classifying these bonds. This process involved examining bond cross-sections through various feature descriptors and classifying them with machine learning classifiers that had been trained on a set of 45 hand-drawn molecules. The promising outcomes achieved with a limited dataset suggest the potential for enhanced accuracy with access to more data.

B. Data Set

The dataset used in our study includes a total of 360 photographs, which feature nine distinct, basic hand-drawn chemical structures. There are 40 photographs for each type of molecule. The drawings were created using a fine point black marker on plain white printing paper. The images were captured using the camera of an iPhone 12 Pro and captures images in three colour channels without an alpha channel. These images were then reduced in resolution to 400 x 300 pixels in grayscale format. Uniform lighting conditions were maintained for all photographs, and three different individuals were involved in the drawing process. This diversity was intended to prevent the model from becoming too narrowly adapted to a single individual's style. The only preprocessing technique employed was the application of binarization at a threshold of 40%. No further preprocessing steps were taken.

C. Text Label Recognition

We investigated a number of text recognition techniques, such as scale-invariant template matching, which was evaluated using five examples of each of six distinct templates (O, H, OR, RO, N, and OH). We also used a number of supervised learning classifiers that used three templates (designated as "O," "H," and "N"). We found that the photos' scale varied from a minimum of 20x20 pixels to a maximum of 60x60 pixels after analysing the dataset. Using a spatial pyramid sliding window technique, we raised each window's size by 5 pixels, going from 20 to 60 pixels. The challenges we had while attempting to use supervised learning classifiers led us to use the scale-invariant template matching approach.

We used a Gaussian filter with a size set to half the stroke width to smooth all of the training templates and the target image. As previously mentioned, we used the spatial pyramid sliding window technique for matching. The maximum F1 score was obtained by identifying the ideal tolerance threshold as 0.77. Our solution to the bounding box overlap problem was non-maximal suppression.

D. Detection of Corners

We have adapted the Douglas-Peucker algorithm for use with our data to gauge its effectiveness against the MLOCSR method. Additionally, we have used a corner detection method that leverages the principles of the Harris corner detector, with some modifications for broader application.

In order to uphold uniformity with MLOCSR terminology, we delineate multiple crucial points: A "T-point" indicates the termination of a line segment at a text box, which suggests a bond with an atom other than carbon.

A "C-point" is a corner at the intersection of a carbon's main bonds. A "D-point" indicates the termination of a line segment that is not connected to the main bond structure, indicating double or triple bonds.

Following the approach of MLOCSR, we utilize the Douglas-Peucker algorithm to pinpoint areas around C-points and T-points, deferring the search for D-points until primary corners have been identified. This method progressively attempts to fit polygons with an increasing number of vertices to each contour, continuing to add vertices until all points on the contour are within a specified threshold distance from the polygon. The vertices of the fitted polygon are then extracted as the output.

To identify clusters within the image, we apply the Canny edge detector and then look for coinciding points across the contours that the polygons fit. We adopt a threshold that is the square root of 2 multiplied by the length of the edge, in line with the guidelines set out in MLOCSR.

Next, we put into practise a straightforward agglomerative clustering technique, capping the spacing between groups at 50 pixels. A polygon vertex is grouped into an existing cluster and the cluster's central point is recalculated if the distance between it and the cluster's centre is smaller than this maximum. A new cluster is formed if it is farther away.

In this case, the goal of using the Harris detector [12] is still to identify C- and T-points without concentrating on the nuances of D-points that distinguish between double and triple bonds. To identify notable alterations in the gradient of the image in two dimensions, the Harris corner detector is employed.

First, the image is subjected to a broad Gaussian filter that corresponds to the predicted width of the strokes. The Harris corner detector is then activated, provided that corners that are identified remain at least as far apart as specified by a preset threshold.

E. Detection of Bonds

The methodology used differs from traditional line vectorization practices commonly referenced in the field, including MLOCSR, by leveraging the Hough transform for the sole purpose of detecting bonds instead of categorizing them. Such conventional vectorization techniques often err, particularly with hand-drawn molecules, where they fail to pinpoint D-points that are identifiable through polygon reconstruction in instances where molecules are rendered with perfect straightness. Considering that advanced methods like MLOCSR have a recognition rate of less than 80% for C- and T-points, it's unlikely that these approaches would accurately detect the more nuanced D-points amid the significant variations found in hand-drawn bonds.

For each node identified previously, we examine the nearest four nodes to check for the presence of a bond, based on the chemical fact that a carbon atom can form a maximum of four bonds. While it's not entirely impossible for additional molecules to attach to a carbon, such instances are exceedingly rare and were not observed in our dataset. Further node analysis for broader chemical structures might involve filtering out false positives using a Markov logic network akin to the one in MLOCSR; however, we haven't included that method in our current methodology to keep things simple.

Another heuristic we apply is the assumption that no bond exists between the two outer nodes when three nodes are linearly aligned, a scenario that typically arises with two bonds at a 180-degree angle.

We obtain an initial list of node pairs that might be bond shares by implementing these conditions. We filter this list by partitioning it into windows of equal sizes and enclosing the edge between two nodes within a bounding box of predefined width (40 pixels). To find lines within each window, we apply a very low threshold Hough transform, taking into account only those that align within one degree of the anticipated bond direction. The next thing we do is make sure that every window in the bounding box has a detected line that matches the Hough transform. We anticipate that at least one window will not align a line in the direction of the node pair if they do not exhibit a bond. This approach produces a low rate of false positives, even if it might cause some false negatives that could be rectified by a Markov logic network in later stages. Markov logic networks would also filter out the rare false positives, which are usually incorrect triangular connections in organic compounds.

F. Classification of Bonds

The final part of the methodology is bond classification. This is achieved by employing a sliding window technique that traverses the bond cross-sections extracted from our training data. The window dimensions are set at 10 pixels in width along the bond's length and 40 pixels in breadth across it. This typically yields between 3 to 10 windows for each bond. Our curated training collection encompasses an assortment of bond types: 62 single, 33 double, 10 wedge, 10 dashed, and 5 triple bonds.

In order to prevent overfitting due to the small size of our training dataset, we avoid using neural networks and instead use Histogram of Oriented Gradients (HOG) features extracted from each sliding window to train different supervised learning classifiers, such as a multiclass logistic regression, a linear Support Vector Machine (SVM), and a decision tree.

We next repeat the procedure for every bond in the test set, extracting HOG features and using the trained classifier to predict the bond type in each window. Next, a voting system is used, in which every window casts a vote to determine which bond type is more common.

IV. RESULTS AND DISCUSSION

A. Text Label Recognition Results

We evaluated the precision and recall metrics on the test dataset in order to optimise the accuracy of the scale-invariant template matching. The tolerance level that produced the highest F1 score, which was discovered to be 0.77, was chosen as the ideal threshold.

We then used our template matching strategy to the test set. Table A shows the performance results, broken down by individual molecule and totalled.. The accuracy was evaluated on a per-image basis, meaning that for a molecule image to be deemed accurately recognized, it had to be completely identified without any false positives. For your reference, the Appendix contains visual representations of the molecules that correspond to each molecule no and chemical name/formula.

B. Detection of Corners Results

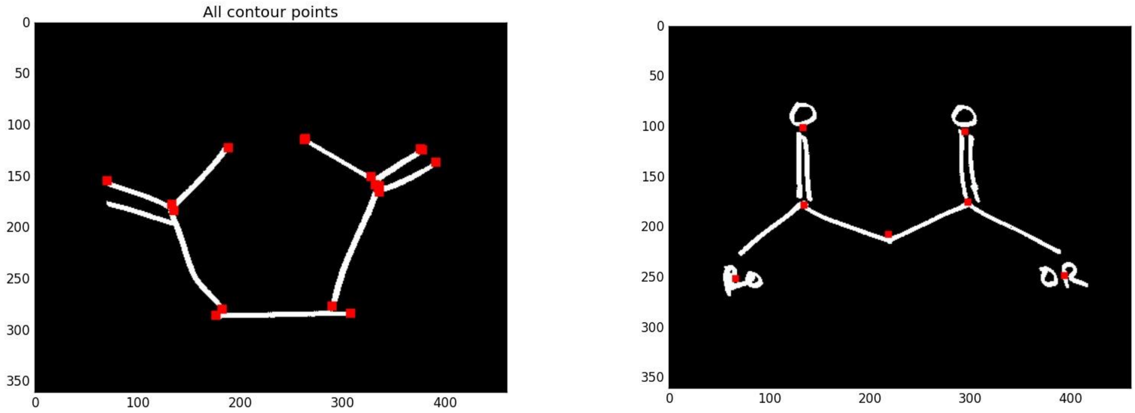
Table 1 clearly demonstrates the superiority of the Harris corner detector over the standard MLOCSR polygon reconstruction technique, with a notable 15% improvement in accuracy at the molecular level. The proportion of correctly identified molecules—that is, molecules without any false positives—to the total number of molecules is what determines this accuracy level. This difference in performance can be attributed to other important factors. Initially, the polygon method does not handle dashed bonds effectively. On the other hand, Gaussian smoothing, which is part of the Harris detector's preprocessing, effectively merges dashed bonds into a single edge for corner detection. This preprocessing step is incompatible with the polygon method, which relies on the distinct opposing outlines that make up the edges of thicker lines. The comparative analysis of dashed molecules, highlights the limitations of both methods in dealing with dashed bonds, with the polygon method achieving only 15% accuracy and the Harris method reaching 45%. This is particularly challenging when dashed bonds, which appear quite broad after blending, are located in close proximity to other corners.

Our predictions are likewise confirmed: the Harris method outperforms the polygon reconstruction in detecting non-linear bonds. This benefit is particularly apparent in the detection of benzene rings, where the polygon approach only reached a 50% accuracy rate, whereas the Harris corner detector achieved a high accuracy rate of 89%.

Table A: Comparison of corner detection methods (Polygon method and Harris Corner method)

	Molecule Accuracy	Overall Precision	Overall Recall
Polygon method	0.50	0.96	0.97
Harris method	0.89	0.98	0.98

Figure A: Douglas Peucker Algorithm Results



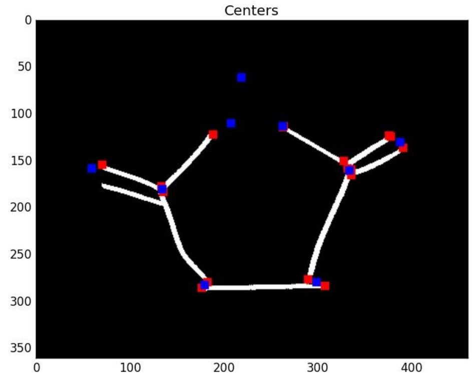


Figure B: Clustering Results

C. Detection of Bond Results

The stage of bond detection is currently the weakest link in the process, yet it also presents the most opportunities for correction through the use of a high-level Markov model, as outlined in the MLOCSR framework. There's significant room for enhancement in the algorithm by integrating more heuristic knowledge regarding atomic bonding patterns. Our current approach does not incorporate chemical insights on molecular topological structures during bond detection, but by applying additional constraints, we could lower the rate of false positives. This would therefore enable us to adjust the angle acceptance criteria and the Hough detector's sensitivity, perhaps lowering the number of false negatives.. Detailed results of these findings are displayed in Table 4, with a more in-depth analysis of the errors shown in Figure B.

Common mistakes involve overlooked bonds, as depicted in Figure C, incorrect formation of triangular structures due to the presence of bonds that are almost in a straight line but not quite, as illustrated in Figure C.

Results of bond detection

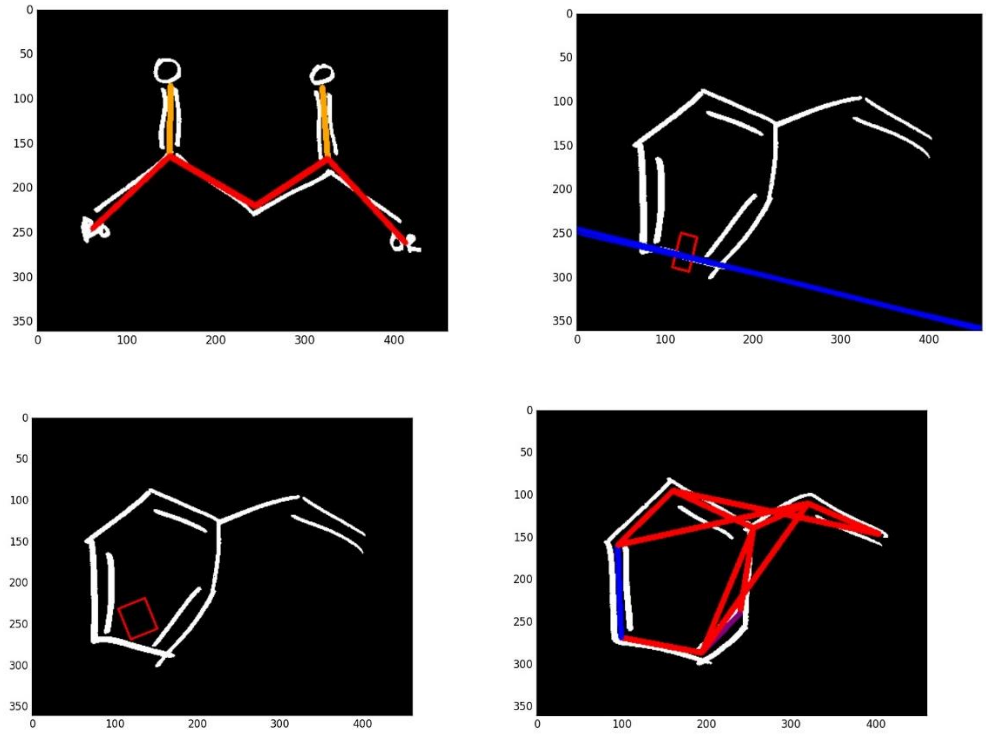


Figure C: Bond detection error likely due to missing bond

Table B: Results of Bond Detection

Molecule ID	Molecule Accuracy	Overall Precision	Overall Recall
1	0.98	0.98	0.98
2	0.29	0.92	0.93
3	0.0	1	0.78
4	0.23	0.97	0.76
5	0.66	0.97	0.88
6	0.79	0.97	0.91
7	0.18	0.82	0.93
8	0.69	0.99	0.89
9	0.07	0.87	0.83
Total	0.43	0.94	0.88

D. Classification of Bonds Results

Comparison of Classifiers: We divided our training data for bonds into two sets, using a 90% portion for training and a 10% portion for validation, and conducted 10-fold cross-validation.

Table C: Cross validation results on 90-10 set

Classifier	Accuracy
Logistic Regression	0.86
Linear Support Vector Machine	0.95
Decision Tree	0.87

Performance on Test Set: Upon analysing the cross-validation data, we decided to implement the Support Vector Machine (SVM) for the classification task across all molecular data. The evaluation reveals a balanced rate of misclassification among different bond types; even with as few as 5 instances of triple bonds in the training data, the likelihood of double bonds being incorrectly identified as single bonds is not higher than them being misclassified as triple bonds.

Table D: Test results using SVM

Molecule No.	**Bond Accuracy**	**Molecule Accuracy**
1	0.96	0.90
2	0.96	0.80
3	0.56	0.0
4	0.79	0.29
5	1.0	1.0
6	0.83	0.50
7	0.92	0.92
8	0.97	0.92
9	0.95	0.86
Total	0.88	0.68

E. Final Results

When the complete processing system is applied to the full batch of molecules, it successfully identifies 94 out of the 360 molecules in their entirety. Although this success rate might not appear high, it exceeds that of pre-existing optical structure recognition software, notably OSRA, which tends to have almost no accuracy with handwritten inputs. Furthermore, the MLOCSR method, which utilizes the Douglas-Peucker algorithm for polygon fitting, is less effective than the algorithm we used at detecting C- and T- points within our handwritten dataset. Remarkably, the bond classification algorithm used, which is based on supervised learning, shows excellent performance despite the limited training set derived from merely 5 images for each type of molecule. Looking ahead, we are confident that by expanding our training dataset, we can achieve close to 100% accuracy using this technique in the future.

V. Conclusions

Despite the modest overall accuracy achieved, the research detailed in this paper is expected to establish a solid groundwork for the recognition of hand-drawn chemical structures in future endeavours. The primary reason for the less-than-ideal accuracy is the scarcity of training data. For instance, adopting cutting-edge OCR technologies could potentially elevate text recognition accuracy from 77% to nearly flawless levels. Moreover, an expansion in training data is likely to enable the transition from an SVM to a convolutional neural network for bond classification, which could significantly refine the precision of this process.

The basic recognition tasks of detecting atoms and bonds—that is, the nodes and edges that make up the molecular graph structure—were the main focus of this work. We anticipate that incorporating higher-order heuristics, such as chemical valence rules which were not considered here, will greatly enhance bond detection capabilities.

It's also worth noting that accuracy might not always be the best measure of performance for hand-drawn structure recognition applications, especially when additional context is available. For instance, digital drawing tools on electronic tablets can provide supplementary data, akin to the methods used by OCR software for recognizing complex characters in languages like Chinese and Japanese. This supplementary information can include the timing of pen strokes, which can improve corner detection, and the stroke speed, which can lead to more accurate bond identification.

Furthermore, the requirement to recognize only a limited range of molecules can allow for the use of molecule similarity algorithms, which compare the drawn structure against a database and match it to the most similar known molecule. This approach could be especially beneficial in educational settings for identifying simple molecules.

In conclusion, the challenge of recognizing and analysing handwritten molecular structures is complex and requires a different approach than that used for computer-generated images. This methodology requires us to take into consideration the different factors that enter the frame due to the dataset being full of hand drawn images instead of digital images.A significant innovation of this project was the decision to analyse small segments of bonds to build a consensus view, rather than attempting to recognize entire bonds at once, which was the norm in previous approaches. While there are numerous aspects of the pipeline that can be enhanced as discussed, substantial strides have been made in adapting these methods for potential public application.

VI. Futurework

As previously mentioned in the Methodology section and the Results section, this project can be further improved with the increase in the size of the dataset used. The training set especially yielded a high accuracy with an SVM classifier with just 5 instances. We are confident that with a better and larger training set, we can achieve even higher accuracy as we will be able to use other machine learning models such as Convolutional Neural Networks which require a large training set. Furthermore, an advanced OCR engine used for text label recognition could also be implemented in order to further increase our accuracy achieved of 77%. All in all, we hope this lays a solid groundwork for future developments in this field of computer vision.

References

[1] Gaulton, A.; Overington, J. P. Role of open chemical data in aiding drug discovery and design. Future Med. Chem. 2010, 2, 903−7.

[2] Kind, T.; Scholz, M.; Fiehn, O. How large is the metabolome? A critical analysis of data exchange practices in chemistry. PLoS One 2009, 4, e5440.

[3] G.R. Rosania, G. Crippen, P. Woolf, D. States, and K. Shedden, R. A cheminformatic toolkit for mining biomedical knowledge. Pharmaceutical Research, vol. 24, (no. 10), pp. 1791-1802, Oct 2007.

[4] Casey, R. et. al. Optical Recognition of Chemical Graphics. Document Analysis and Recognition: Proceedings of the 2nd International Conference on Document Analysis and Recognition. 1993.

[5] Ibison, P. et. al. Chemical Literature Data Extraction: The CLiDE project. Journal of Chemical Informatics and Computer Science, 33, pp. 338-344. 1993. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[6] Park, J. et. al. Image-to-Structure Task by ChemReader. Text Retrieval Conference, 2011.*FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[7] Filippov, I. and Marc Nicklaus. Optical Structure Recognition Software to Recover Chemical Information: OSRA, An Open Source Solution. J. Chem. Inf. Model., 49 (3), pp. 740-743, 2009.

[8] Frasconi, P. et. al. Markov Logic Networks for Optical Chemical Structure Recognition. Journal of Chemical Information and Modeling. 54, pp. 2380-2390, 2014.J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[9] Douglas, D.; Peucker, T. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can. Cartogr. 1973, 10, 112−122.

[10] Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR '05), Jun 2005, San Diego, United States. IEEE Computer Society, 1, 886–893, 2005.

[11] de Campos, T.E.; Babu, B.R.; Varma, M. Character Recognition in natural images. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, February 2009.

[12] Harris, C.; Stephens, M. A Combined Corner and Edge Detector. Plessey Research, 1988.

[13] Rajan, K.; Brinkhaus, H.O.; Zielesny, A.; Steinbeck, C. A review of optical chemical structure recognition tools. J. Cheminformatics, vol. 12, 60, 2020. https://doi.org/10.1186/s13321−020−00465−0.

[14] Rajan, K.; Brinkhaus, H.O.; Agea, M.I.; Zielesny, A.; Steinbeck, C. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific

publications. Nature Communications, vol. 14, 5045, 2023. https://doi.org/10.1038/s41467−023−40782−0.

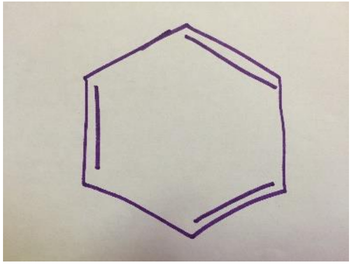
[15] Filippov, I.V.; Nicklaus, M.C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. J. Chem. Inf. Model., vol. 49, 740–743, 2009. https://doi.org/10.1021/ci800067r.

[16] Hormazabal, R.; Park, C.; Lee, S.; Han, S.; Jo, Y.; Lee, J.; Jo, A.; Kim, S.; Choo, J.; Lee, M. CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition. In 36th Conference on Neural Information Processing Systems (NeurIPS 2022), Track on Datasets and Benchmarks, 2022.

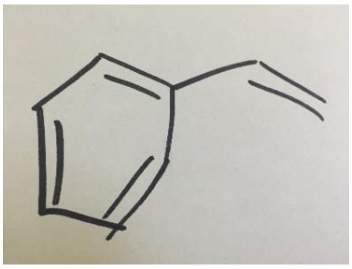
APPENDIX

Molecule No. - Chemical Name/Formula

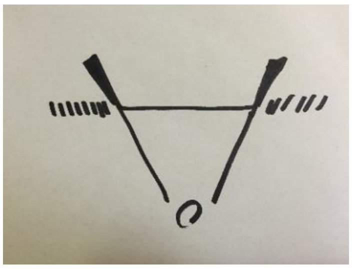
1. Benzene – C6H6



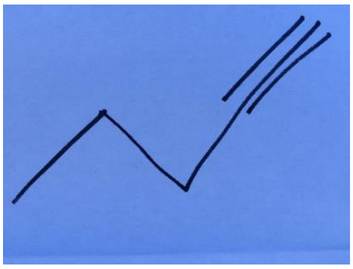
2. Styrene – C8H8



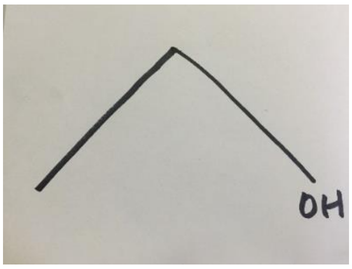
3. Hexanal – C6H12O



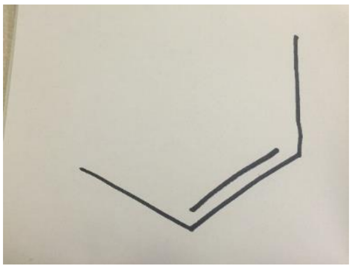
4. **(**1 – Pentyne**)**– C5H8



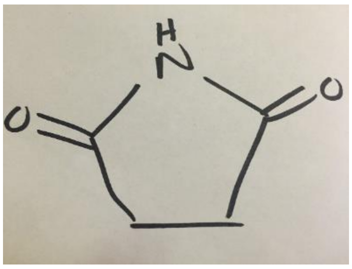
5. Ethanol – CH3CH2OH



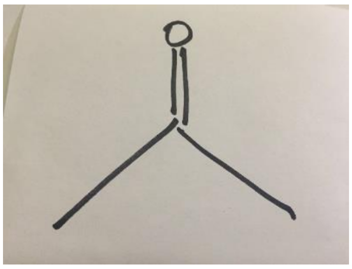
6. (cis – 2 – Butene) – C4H8



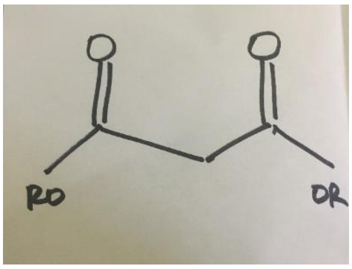
7. Succinimide – C4H5NO2



8. Acetone – C3H6O



9. Glyceric Acid – C3H6O4



CONTRIBUTION

A. Nehal Murdeshwar

For this project, we immediately realised that a lot of work had to be done, both on understanding the concepts required and the working behind existing OCSR technologies and execution of the same concept but on hand-drawn structures which is a lot more difficult. In order to get the other 2 members up to scratch , I researched the topic extensively, looking at multiple research papers in order to create a timeline of events that we had to follow in order to finish the project on time. I, then discussed all of this with my team members and we together came up with a distribution of work.

My contribution mainly was to help in the coding of the project. In our pipeline, we have 6 different tasks that are all interrelated. I began by implementing the first 2 tasks which were scale invariant template matching, text removal along with various preprocessing elements required. The main gist of the project was in the detection of corners and bonds section on which I helped my other 2 team members, but they were the ones that were front lining most of the coding process with suggestions from me. Along with that, I contributed to the making of the report, covering everything up and till the literature review section. Finally, I refurbished the report at the end making some slight changes to formatting issues, and the general structure of the report.

B. Jayasuryan Mutyala

I worked on implementing the bond detection algorithm and collaborated with Abhiram on the classifier implementation. I was responsible for developing the detect_bonds function, which maps out the molecular bonds by combining the retrieved data points and generating a network representation of the chemical structure. This function makes the verification process easier by not only identifying the bonds but also visualizing the connections.

In addition, I implemented the corner_detector function, which highlights and identifies corner features in molecular images using computer vision techniques including canny edge detection, harris corner detection, and others. By preparing the data for additional analysis, this function establishes the foundation for the extraction of structural information from molecular images.I also worked on the function that detects bonds between corners. It uses Hough transform techniques to establish the existence of linear features that indicate bonds and analyses spatial linkages and angular consistencies within the image to discover the presence of chemical bonds between selected corner points. In the report I worked in writing methodology.

I made sure the image processing pipeline was reliable, accurate, and efficient by contributing these changes to the project, which improved our machine learning classifiers' capacity to decipher complex molecular pictures.

C. Abhiram Reddy

In order to categorise chemical bonds from image data, I concentrated on creating the fundamental image processing and machine learning pipeline. My main contribution was to the preprocess_training function, where I designed a technique to create consistency for machine learning model training by normalising the widths of chemical bond pictures.

Additionally, I put into practise a unique feature extraction method called Histogram of Oriented Gradients (HOG), which is contained in the hog function. Through this procedure, the visual input was converted into a format that our classification systems could handle with ease. In addition, I worked on the get_bonds function, which prepared sub-images of possible bonds for classification by algorithmically identifying and extracting them from a larger image. For the analysis of intricate molecular structures, this was essential. For contritbutions to the write up for the report I wrote contributed to the result analysis.

Ultimately, the bond classification pipeline was completed by my classify_bonds function, which used the trained classifier to assign bond types to each detected bond in the image and extract insights. Additionally, this function integrated user feedback to enhance the model.