CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition

Aniko T. Valko and A. Peter Johnson*

Keymodule Ltd., Hobberley Lodge, Hobberley Lane, Leeds LS17 8JQ, United Kingdom

Received December 12, 2008

We present CLiDE Pro, the latest version of the output of the long-term CLiDE project for the development of tools for automatic extraction of chemical information from the literature. CLiDE Pro is concerned with the extraction of chemical structure and generic structure information from electronic images of chemical molecules available online as well as pages of scanned chemical documents. The information is extracted in three phases, first the image is segmented into text and graphical regions, then graphical regions are analyzed and where possible the connection tables are reconstructed, and finally any generic structures are interpreted by matching R-groups found in structure diagrams with the ones located in the text. The program has been tested on a large set of images of chemical structures originating from various sources. The results demonstrate good performance in the reconstruction of connection tables with few errors in the interpretation of the individual drawing features found in the structure diagrams. This full test set is presented for use in the validation of other similar systems.

1. INTRODUCTION

Chemists communicate structural information for compounds via depictions of two-dimensional chemical structures (structure diagrams). Nowadays, these are usually created using chemical drawing programs that capture complete structural information. However, nearly all printed and electronic sources of the chemical literature including reports, journals, and patents present structure diagrams in the form of images. Although chemical images can be easily interpreted by the human expert, they lack the explicit structural information required for input to chemical databases or chemistry software applications. The reproduction of this information by redrawing the structure with a software tool is time-consuming and liable to errors but nonetheless is still the norm for these purposes. Thus there is a pressing need for an efficient optical chemical structure recognition or OCSR system that can automatically turn digitized structure diagrams into structure descriptions − connection tables or equivalent line notations − which are suitable for input into chemical structure databases.

Interest in OCSR dates back to the beginning of the 1990s when four projects − Kekulé,[1,2] the Contreras system,[3] the IBM system,[4] and CLiDE (Chemical Literature Data Extraction)[5] − were developed and in some cases evolved into commercial products. In recent years, new OCSR tools have emerged, including chemOCR[6−8] and two free open-source programs, ChemReader[9] and OSRA.[10] Some of these systems have novel features and validation studies show some promise, but nevertheless we believe that none of the current systems represents a complete solution to the OCSR problem.

Because of this unmet need, we have developed a new version of the long-term CLiDE software, called CLiDE Pro. The primary aim of CLiDE Pro is to overcome some of the limitations of CLiDE and thus achieve a good recognition performance on a diverse set of structure diagrams and also to include additional recognition capabilities, for example in the area of patents. In this paper, we present features of the current system, including a partial solution for generic structures as well as the methods for dealing with a variety of difficult drawing features. CLiDE Pro's accuracy is demonstrated via an extended and detailed validation, and its current limitations and future development are discussed.

2. SYSTEM DESCRIPTION

Since the vast majority of published articles are available for download in PDF (Portable Document Format) file format, it was essential that CLiDE Pro could accept these files as one form of primary input. The system supports document-oriented processing of PDF documents as opposed to page-oriented processing, in the sense that a whole document can be processed at the same time, which makes it possible to link together related information located on two or more document pages. However, the user can specify a subset of the total set of pages or even select individual structures to be subjected to interpretation. In addition to pdf documents, images of chemical molecules (tiff, bmp, etc.) and of whole journal pages can also be loaded into the system. The output produced by CLiDE Pro is provided in its own feature rich CLiDE format as well as the MDL Molfile format which is readable by a wide range of chemical software tools.

There are three main problems involved in OCSR: (1) identification of chemical images within a document, (2) compilation of chemical graphs of individual molecules from chemical images, and (3) interpretation of complex objects such as generic structures using the retrieved chemical

* Corresponding author e-mail: Peter.Johnson@keymodule.co.uk.

10.1021/ci800449t CCC: $40.75
Published on Web 03/19/2009

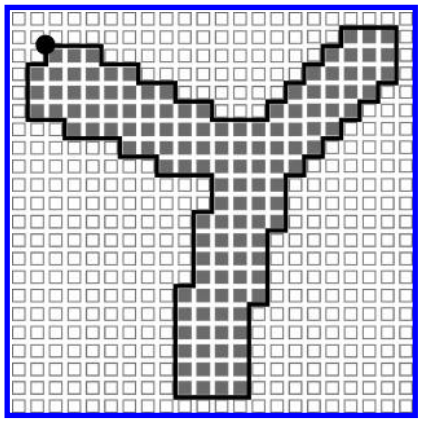

Figure 1. The interpixel contour of a connected component represented with a series of directions: E,E,E,S,E,E,S,E,E,S, etc.

graphs. In the following sections, our solutions to these problems are detailed.

2.1. Identification of Chemical Images. As structure diagrams are usually drawn with black ink on a white background, earlier systems were often limited to dealing with bilevel images consisting of *on*-pixels (black) and *off*-pixels (white). However, structure drawings may contain colors other than black for various reasons, e.g. for indicating atom types. Additionally, the color of the background can be different from white. In order to separate the individual elements consisting of a molecule in a color-scaled image from the background, the input image is binarized. During *binarization*,[11] a color-scaled image is turned into a bilevel image by classifying every pixel as an *on*-pixel or as an *off*-pixel based on a threshold-based binarization technique. The image is then segmented into connected *on*-pixel regions or *connected components*, based on a nonrecursive technique which scans the image on a row basis and identifies sets of adjacent horizontal *on*-pixel segments. A connected component is then represented with its interpixel contours which are defined as the coordinates of a starting point and a sequence of four directions (North, South, East, West) following the contour of the connected component surrounding its *on*-pixel region (Figure 1).

The digitized image of a document page consists of a mixture of text and graphics. To automatically identify graphic regions, CLiDE Pro performs *document image segmentation* based on a method[12] which builds up the tree structure of a page in a bottom-up manner, i.e. it starts by processing the connected components of the image, and results successively in a list of words, text lines, text, and graphic blocks. The layout analysis first calculates the distances between enclosing boxes around individual connected components. The connected components are considered as the vertices of a graph, and the distances between them as the weighted edges of the graph. Words, lines, and blocks are then derived from the minimal-cost spanning tree, built with Kruskal algorithm.[13]

2.2. Reconstruction of Molecular Structures. The connection table of a digitized chemical structure is established in five phases: (1) classification of connected components into basic groups, (2) vectorization that converts graphical connected components into graphs of vectors, (3) construction of dash bonds from connected components classified as dashes, (4) construction of atom labels, and (5) compilation of connection table. The individual phases are explained in the following paragraphs.

Classification of Connected Components. Connected components are first divided into basic subclasses such as characters, dashes, lines, graphics, and noise. A stepwise procedure is implemented where in each successive step the elements of one group are separated from the others. This process uses statistical data extracted from the image based on a method[14] which analyzes the distribution of connected component dimensions and estimates a size threshold value, for which all graphic components are expected to be larger than this threshold value and all other components are smaller. The classification of connected components is performed using this threshold value together with other simple features such as the relative height/width, the on-pixel density, and the number of contours.

Vectorization. The connected components of the lines and graphics subclasses are vectorized to produce a set of vectors representing bond lines. During this process, the contours of the connected components are analyzed, and an approximation polygon is created for each contour by identifying straight and curved contour fractions based on a method similar to Sklansky and Gonzalez.[14,15] This polygon is created in such a way that each point of the original contour is within a threshold distance value of a side of the polygon. If this threshold value is well chosen, straight parts of the contour result in long polygon sides, while curved parts are approximated by consecutive short sides. Long polygon sides are chosen as straight contour fractions, and consecutive short sides are merged into curved fractions. Individual short sides are ignored henceforth. Vectors are then found by searching pairs of adjacent fractions. Figure 2 illustrates the individual steps performed during the vectorization process. In an ideal case, there are two fractions created for each linelike object, which are the two borders. However, due to noise or some other error, a border of a line could be cut and detected as more than one fraction. Alternatively, borders could be left uncut, or the fraction created could be shorter than the actual border of the line. These errors are recognized and corrected by cutting and joining fractions using information about the fraction at the opposite border (Figure 3). Finally, each vector is described by the two borders. The coordinates of the line end points, the line width, and the shape, such as straight or wedged, are determined from these two borders. Following vectorization, wavy lines are perceived by searching for groups of vectors that are not longer than the average character size, touch each other at their end points, and lie on a straight line.

Construction of Dashed Bonds. Dashed-line detection is based on the Hough transform method.[16,17] This is done by searching for sets of connected components which have been classified as dashes, situated along a straight line. Wedged dashed lines are perceived on the basis of the size difference of dashes located at the two ends of lines. Straight dashed lines are then further analyzed to recognize whether they represent lengthwise-oriented (collinear dashes) or crosswise-oriented (dashes parallel to each other) dashed lines. After this process, dashes not belonging to any detected dashed line are reclassified as characters, lines, or noise according to their size and location compared to their surrounding connected components.

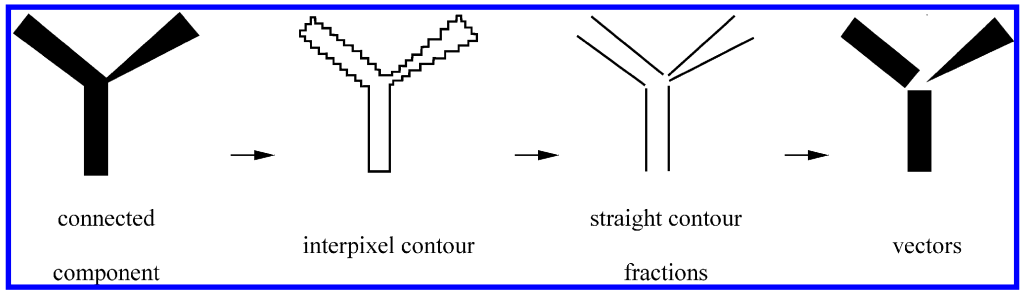

Figure 2. Steps of vectorization.

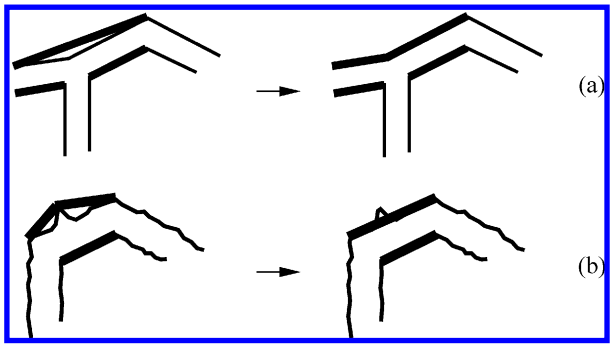

Figure 3. Two kinds of fraction correction performed during vectorization: (a) cutting of a fraction into two and (b) merging of two fractions into one.

Construction of Atom Labels. Atom labels are constructed in three steps. First, using an OCR engine, individual characters are interpreted based on a topological and geometrical feature analysis and a filter technique for the classification of characters. Apart from tiny characters, the OCR engine can interpret symbols of different size and font and is tolerant to some degree of rotation. Recognized characters are then grouped to form words based on their coordinates. This process not only combines characters lying next to each other but also considers vertical relationships to reconstruct vertically aligned atom labels. Finally, atom labels are identified in accordance with a superatom database which includes all the elements of the periodic table, functional groups which occur frequently in structure diagrams, and labels commonly used to represent R-groups in generic structures such as *R*, R_1, *R*′, and *X*. For every item contained therein, the information held in the database includes the name, the nature, the atomic number for atoms, and the connection table for groups. Longer atom labels, i.e. linear representations of structural formulas not found in the database (e.g., CH_2CH_2OH), are parsed.[12]

Reconstruction of Molecules. Connection tables are established from the recognized solid and dashed lines and atom labels.[18] To determine the connections among atoms and bonds, a bond line is represented with its two end points and a free-flag associated with each end. The free-flag is set to false if a bond line touches another bond line with the end to which the free-flag belongs; otherwise it is true. First, it is determined which bond lines are connected to the atoms. A bond line end is joined to an atom if (i) the distance between them is smaller than a certain threshold, (ii) the free-flag associated with the bond end is true, (iii) the bond points toward the atom label, and (iv) the bond end is not closer to any other atoms. Bond ends not connected to any atom label are then joined together to form implicit carbon atoms. The primary criterion for joining two bond ends is that they lie sufficiently close to each other. If the two bond ends belong to the same connected component, then it is also required that either the two bond ends have the same free-flag value or the two bonds are parallel. After all connections among the bonds and the atoms are found, the connection table is built where two or three bond lines connecting the same atoms, including implicit carbon atoms, are merged to form double and triple bonds. For most of the bonds, the sense of direction of the bonds is not relevant, but for the wedged bonds and dashed wedged bonds, the connection table can distinguish which atom is located at the apex of the wedge and which atom is located at the opposite end.

2.3. Interpretation of Generic Structures. A generic substance represents more than one substance or a set of specific substances. This set of substances can be represented by a generic structure which usually comprises an invariant part together with associated variable groups or R-groups. In most cases, the structure lies inside a graphical region, and the R-group labels along with the substitution values lie inside a text region in which R-groups are often listed sequentially and the individual substitution assignment expressions have the following form: variable, 'equals' sign, and one or more substituents separated with 'or' logical tokens. R-group assignments can be expressed in formats other than linear text format such as tables of R-groups and substitution values without the presence of the 'equals' sign or partial structure diagrams defining the substituents. Figure 4 shows examples of the different formats which are widely used in the literature to describe R-group assignments.

In CLiDE Pro, the interpretation of generic structures is performed in two phases. In the first phase, the generic text blocks are identified and separated from other text blocks and are interpreted using a generic text interpreter whose task is to extract the generic information from the text. The stages involved in this task include determining the R-groups, the number and type of the substituents, and whether any label is present for each substituent, etc. The generic text interpretation[12] is performed in three steps: (a) lexical analysis or tokenization which isolates the individual words (or tokens) of the sentences, (b) syntax analysis which detects the parts of the sentence requiring contextual checking, and (c) semantic analysis which determines the meaning of the words. The current generic text interpreter is limited in that it handles only linear text format and relies on the presence of special symbols such as an 'equals' sign separating R-groups and substituents and delimiters such as commas and semicolons. Extension of the generic text interpreter to interpret R-groups assignments expressed in various formats is under way.

In the second phase, the most suitable generic text block is identified for a given structure in two stages. First, a search is performed to find the generic text blocks which best match

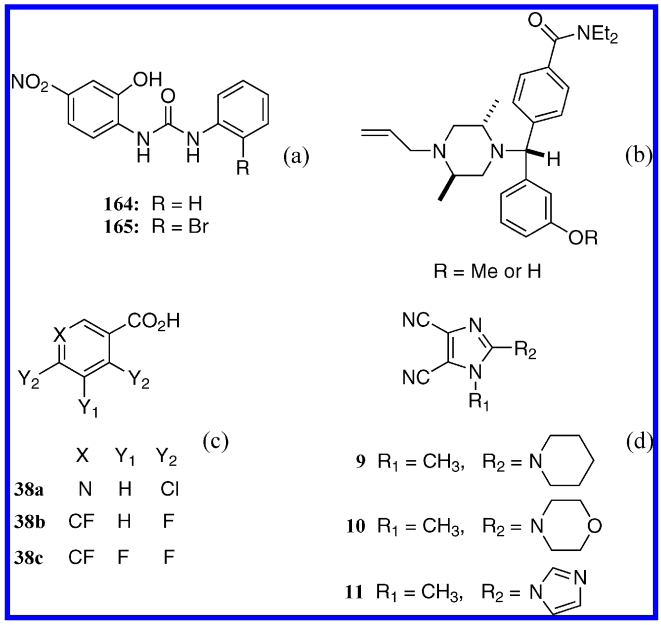

Figure 4. Generic structures having their R-groups and substituents defined in various formats: (a) linear text format with single substituent assignments, (b) linear text format containing a multiple substituent assignment expressed in 'or' logic, (c) table format, and (d) linear format with substituent assignments containing substitution values defined by partial structure diagrams.

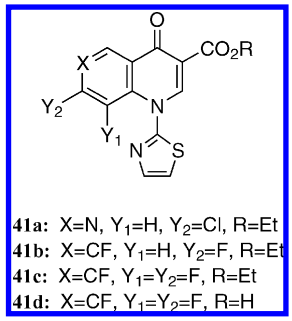

Figure 5. A generic structure interpretable by CLiDE Pro.

the structure in terms of the number of R-groups present in both the structure and the generic text block. Then the generic text block lying closest to the structure is selected from those found in the first stage. A match between a structure and a generic text block goes beyond the check to see whether an R-group in a structure matches an R-group in the text. For instance, as shown in Figure 5, an atom label might be CO_2R in the structure, whereas the R-group in the text might be *R*, which is the part of CO_2R. A simple comparison between *R* and CO_2R is not enough, instead it is the generic element in a generic type which is important. The generic element in both CO_2R and *R* is *R*. Therefore 'match' means that the generic elements in the generic atom labels of the structure and the generic elements in the generic text block are the same.

Figure 5 depicts a generic structure which CLiDE Pro interprets successfully by identifying the R-groups *X*, Y_1, Y_2, and *R* in the text, recognizing the R-group substitution values *CF*, *Cl*, *Et*, *F*, *H*, and *N*, and detecting the generic elements in the text and in the generic atom labels of the structure (*X*, Y_1, Y_2, CO_2R) to find a match between the text and the structure. CLiDE Pro is also able to handle combined R-group assignments such as those in the last two rows of the generic text where Y_1 and Y_2 are assigned to the same substituent, *F*, in one combined assignment.

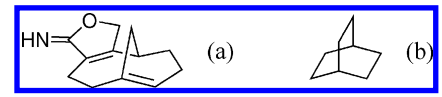

Figure 6. Bridged structures containing crossing bonds (a) with a gap and (b) without a gap at the crossing point.

3. DIFFICULT DRAWING FEATURES

In OCSR it is very likely that an incorrect connection table is generated if there are no specific rules to detect that a structure diagram contains a feature which is unusual or conveys an ambiguous situation. In order to deal with a number of difficult drawing features, a set of rules is used at various stages of the recognition process.

Crossing bonds represent one of the most difficult drawing features of chemical drawings. They are often used to preserve some sense of the three-dimensional shape of the molecule in the drawing, especially in bridged structures. In a crossing bond situation, one bond usually 'cuts' another, although there are many cases where there is no apparent gap in the bond that is being crossed. Figure 6 depicts both types of crossing bond situations. Without special treatment of the crossing bonds, the point where the bonds cross each other can be easily misinterpreted as being a carbon atom. For this reason, a set of rules has been implemented in CLiDE Pro to correctly detect and interpret different crossing bond situations. These rules include the proximity, length, collinearity, and ring membership of potential crossing bonds.[19]

It is also difficult to handle connected components which cause *ambiguity in interpretation*. This mainly concerns certain types of single lines or circles. For instance, a vertical line can occur in several different kinds of chemical entities such as single and multiple bonds, dashed bonds, and character strings representing atom labels (e.g., *Cl*, *I*). Similarly, a horizontal line can be part of a single or a multiple bond, or a dashed bond, or it can represent a negative charge for an atom. In addition to representing oxygen atoms circles are also often used to represent aromaticity in benzene rings, especially in older documents, as well as circles around chemical charges. Such ambiguous situations are resolved by analyzing the environment of the connected components and applying a collection of rules with conditions on chemical and spatial context. For instance, if a 'C' letter is on the left side of a vertical line which is not part of a dashed bond, the vertical line represents the 'l' letter of a chlorine atom. Alternatively, if a circle is located inside a ring with its center near the center point of the surrounding box of the ring and the circle is not too small relative to the ring's enclosing box, then the circle represents aromaticity in the ring encircling the circle. Figure 7 illustrates structure diagrams containing different drawing features that could cause ambiguity in interpretation without special treatment but are correctly treated by CLiDE Pro.

Some bond formations can be easily misinterpreted, and postprocessing of the interpreted connection table is needed to obtain the correct result. For the single bond and a triple bond joined together in the way shown in Figure 8a, use of just the end points of the vectors calculated for each bond line would lead to recognition as a long single bond and a

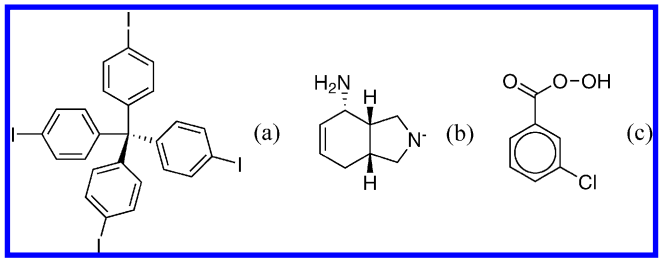

Figure 7. Ambiguity in interpretation illustrated with structure containing: (a) vertical lines representing iodine atoms and a bond line forming a double bond in the upper benzene ring, (b) short horizontal lines representing dashes of a dashed bond and a negative charge, and (c) circles representing oxygen atoms as well as aromaticity in a benzene ring.

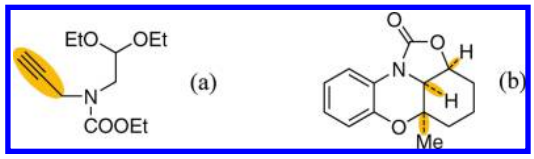

Figure 8. Structures illustrating bond formations which can be easily misinterpreted: (a) a joint single and triple bond and (b) broken-line bonds attached to implicit carbon atoms.

Table 1. Statistical Analysis on the Accuracy of CLiDE Pro's OCSR Method in Regards to the Images and Structure Diagrams Used in the Validation

objects	total number	no. of objects containing errors	success rate
images	454	52	88.55%
structure diagrams	519	53	89.79%

double bond halfway over the single bond. This situation is recognized and dealt with correctly. Figure 8b shows another bond formation requiring special treatment. Here, a broken line directly attached to an implicit carbon atom of a ring can be misinterpreted as a broken-line dashed bond and a short single bond with a carbon atom between the two bonds. Therefore, the average bond length in a constructed connection table is calculated in CLiDE Pro and used to identify short and long bonds, and nearby bonds are then analyzed to see if a correction to these bonds is needed.

4. RESULTS

A systematic test of the OCSR method implemented in CLiDE Pro has been carried out on 454 images containing a total of 519 structure diagrams. All the images originate from published materials including old ones where scanned images were used. In the case of PDF documents, individual images embedded in PDF files are autoextracted in their original full resolution, and those images depicting chemical structure diagrams are selected for inclusion into the test set. Artificial material such as structure diagrams directly produced by chemical drawing software, and hence lacking real artifacts such as noise, has not been included in our test set. The test was performed in a fully automated manner without any human intervention. A high level of accuracy is indicated by the retrieval rate of 89.79% with 466 molecules correctly reconstructed (Table 1). Table 2 analyzes the recognition performance on the individual drawing features present in the test images.

Atom Labels. The biggest proportion of errors occurred on atom labels, with 58 atom labels wrongly interpreted. Atom labels consisting of more than one character proved to be the most problematic in that 44 of them contain errors. As shown in Table 3, the errors mainly originated from the incorrect grouping of characters into atom labels by either breaking single atom labels into parts or merging close but separate atom labels into one. Further reasons for the errors are joined characters and connected components − representing drawing features such as bonds, dashes, and atom index − that touch or lie close to atom labels. In such situations, bad OCR results are produced, and incorrect inclusion of the nearby connected components into atom labels is performed. In one test example, a short single bond is attached to two atom labels, resulting in one merged atom label including the two atom labels and the bond. Table 3 also shows that 5 atom labels are lost because they are attached to the bonds located in their vicinity. Each of the five atom labels consists of just one character, and the characters are vectorized and turned into small bonds. These examples demonstrate that atom labels attached to bond lines frequently result in incorrect detection of atoms and bonds.

Table 2 indicates that two positive charges have been lost during the test. According to the debug messages generated on these two examples, the '+' signs are correctly interpreted by the OCR engine and detected to be formal charges but are lost in the atom label construction phase because they are not associated with their atoms.

Although CLiDE Pro has failed to reconstruct 58 atom labels, some of which contain more than error, a very high level of accuracy of the atom label construction is indicated in Table 2 by a success rate of 98.54% reached on 3981 atom labels. Moreover, all the chlorine atoms (which occur quite frequently in the test images), the iodine atoms, and the negative charges have been correctly detected and interpreted, which was not the case with earlier versions of CLiDE.

Straight Solid Bonds. As detailed in Table 2, the tested structure diagrams contain a large number of straight solid bonds including 16074 single, double, triple bonds, and wedged-styled single bonds altogether. These bonds consist of 20098 straight bond lines as a whole. The error rate for the recognition of straight solid bonds is extremely low, with only 13 bonds wrongly interpreted.

As shown in Table 4, ten bond lines are lost because they are either used in the reconstruction of atom labels as discussed above or not vectorized properly. The vectorization process has failed for the following reasons: (1) a very short bond attached to a longer bond is lost because the vectorization method ignores this short bond, (2) two adjacent bonds with an angle of about 150° between them are vectorized into one bond, (3) the generated vector is much shorter than the bond, and (4) no vector is generated along a bond. Table 4 also indicates that a double bond highlighted in Figure 9 is not interpreted correctly. Although both bond lines forming the double bond are vectorized properly, two single bonds are generated which are connected to the same atom at one end and attached to two separate atoms at their other ends. Finally, a single bond has failed to be connected to its neighboring implicit carbon atom.

Dashed Bonds. Test results reported in Table 2 show a high success rate of 98.37% reached on the reconstruction of 308 crosswise-oriented and two lengthwise-oriented dashed bonds. Although all dashes (except for one which

Table 2. Comprehensive Statistical Analysis of the Accuracy of CLiDE Pro's OCSR Method Measured on the Various Drawing Features Found among the Test Images

drawing features	total number	no. of drawing features containing errors	success rate
atom labels	3981	58	98.54%
horizontally aligned atom labels containing more than one character	1340	41	96.94%
vertically aligned atom labels containing more than one character	272	3	98.90%
chlorine atoms	183	0	100.00%
iodine atoms	8	0	100.00%
negative charges	7	0	100.00%
positive charges	15	2	86.67%
solid straight bonds (including wedged bonds)	16074	13	99.92%
bond lines forming solid straight bonds	20098	11	99.94%
dashed bonds	308	5	98.37%
dashes forming dashed bonds	1333	20	98.50%
wavy bonds	14	4	71.42%
bond crossing situations	28	1	96.43%
aromatic ring circles	21	1	95.23%

Table 3. Error Analysis on Atom Labels

error type	occurrence
an atom label is split into two or more atom labels	24
two or more close atom labels are merged into one atom label	3
bad OCR (including characters touching each other)	8
a close or touching bond is added to atom label	4
a close dash is added to atom label	2
a close atom index is added to atom label	2
an atom label is lost because it touches the adjacent bonds and vectorized to form short bonds	5
an atom charge is lost by not associating to its atom	2
a terminal atom is lost because the neighboring bond is not formed	2
a nearby label is turned into an atom label and joined to the structure	1

Table 4. Error Analysis on Solid Straight Bonds

error type	occurrence
a bond line is lost (added to atom labels)	4
a bond line is lost (vectorization failure)	6
a double bond is turned into two single bonds	1
the connection of a single bond to an implicit carbon atom is failed	1

touches a single bond and was therefore disregarded during the construction of dashed bonds) are correctly identified during the initial classification of connected components, Table 5 reveals that two dashed bonds are fully lost and another is only partially reconstructed. Each of the two lost

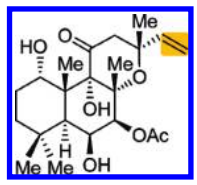

Figure 9. A structure diagram in which a double bond is interpreted by CLiDE Pro as two single bonds.

Table 5. Error Analysis on Dashed Bonds

error type	occurrence
the construction of a dashed bond is failed	2
a dashed bond is partially constructed	1
a terminal dash touching a single bond is not added to the constructed dashed bond (turned into a single bond)	1
a terminal dash touching an atom label is not added to the constructed dashed bond	1
a lengthwise-oriented dashed bond is turned into two single bonds	1

dashed bonds is connected to a terminal atom in the original structure diagram. Hence, the loss of these bonds entails the loss of the terminal atoms. Furthermore, a lengthwise-oriented dashed bond is interpreted as two single bonds.

Wavy Bonds. Table 2 reveals that 4 out of the 14 wavy bonds are not interpreted correctly. The retrieval rate of 71.42% which is noticeably lower than the success rates reported for other drawing features. This reflects the fact that wavy bonds are complex objects requiring advanced techniques to detect the individual waves and group them to form wavy bonds. Figure 10 depicts a structure with two touching wavy bonds which cause the following errors in the

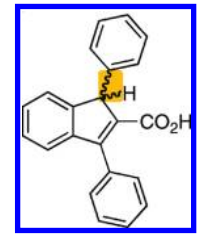

Figure 10. A structure diagram which CLiDE Pro interprets incorrectly due to the two overlapping wavy bonds.

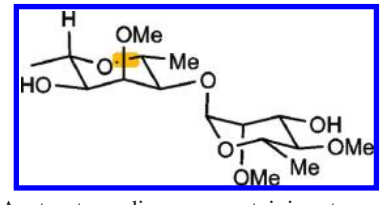

Figure 11. A structure diagram containing two crossing bond situations. CLiDE Pro interprets the crossing bond in the right ring correctly, while the crossing bond in the left ring is not formed correctly because of the shortness of one of the two halves of the crossed bond.

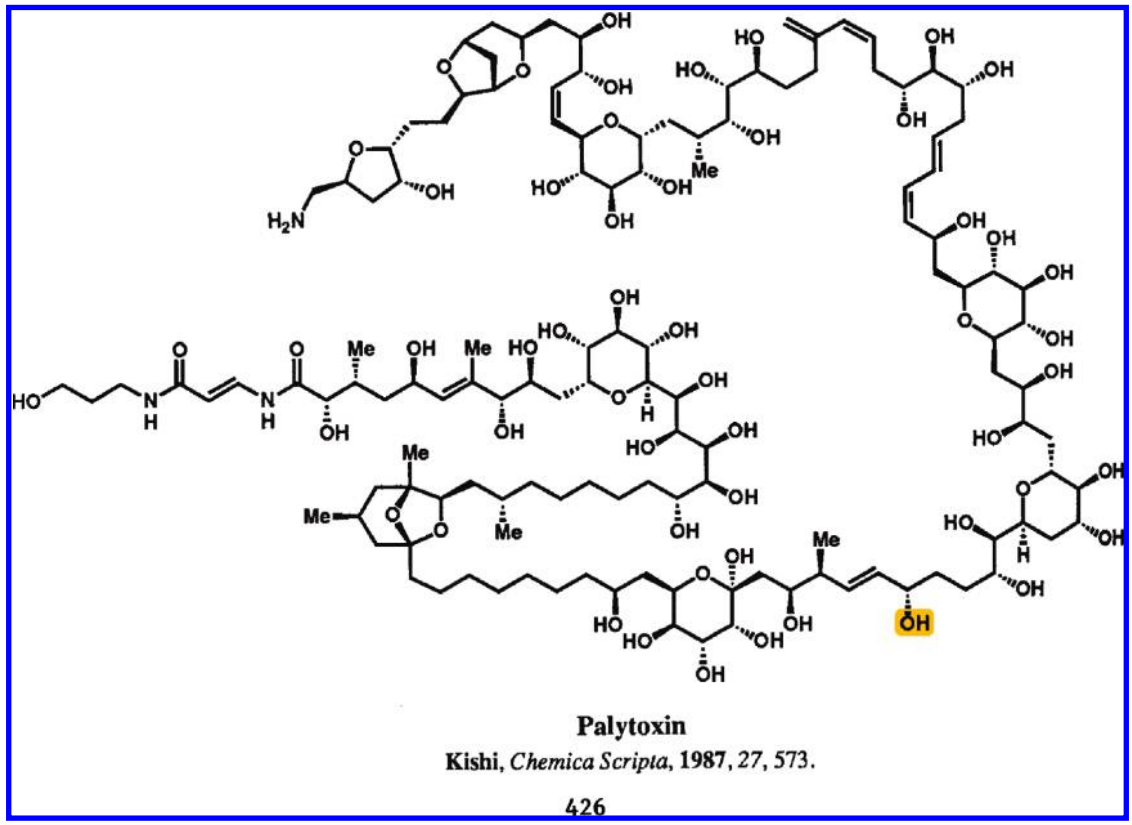

Figure 12. The structure diagram of palytoxin which is processed by CLiDE Pro within a few seconds producing one OCR error in the 'OH' atom label that is located in the lower right part of the diagram and is touched by a dash at the top of the letter 'O'.

reconstruction: the wavy bond connecting the hydrogen atom to the 5-membered ring is interpreted as two chained single bonds, and the wavy bond connecting the two rings is not formed breaking the reconstructed structure into two molecules. We consider this structure a difficult example because the two wavy bonds that overlap one another at their ends require a sophisticated mechanism to identify the waves − especially those that touch each other − and separate them into two groups to form the two wavy bonds. Fortunately, such attached wavy bonds do not occur frequently in the chemical literature.

Crossing Bonds. With respect to 28 crossing bond situations found in our test, the only one not recognized correctly is illustrated in Figure 11 and occurs because one of the two pieces of a crossed bond is very short consisting of only a few pixels and the two halves of the crossed bond are not merged together to form a single bond.

Aromatic Ring Circles. The test images contain 21 circles representing aromaticity in benzene rings. All of the aromatic ring circles are identified, except for one where the shape of this object was not recognized as a circle.

Palytoxin, a Complex Structure. Figure 12 depicts palytoxin, a complex structure containing several atoms and bonds of different styles indicating stereochemistry. This image of palytoxin originates from an old document in which images of structure diagrams are digitized via scanning, which introduces noise in the form of black spots. CLiDE Pro processes this structure within a few seconds by successfully eliminating all the isolated small black spots and overcoming all the distortions in the shape of connected components caused by black spots touching individual connected components. Only one atom label is not interpreted correctly because the 'O' letter of an 'OH' atom label located in the lower right part of the image is merged with a dash of a dashed bond, and therefore the 'O' letter is misinterpreted by the OCR engine. Nevertheless, the dashed bond is formed and connected to an unspecified atom label. In a fully automatic execution mode, unrecognized atom labels are marked in the exported MDL Molfile so that they can be identified quickly for correction. In interactive mode, the program prompts the user with the unrecognized atom label which either can be left unspecified, postponing the correction until the whole interpretation process has finished, or can be corrected on the fly, resulting in a correctly reconstructed connection table.

5. CONCLUSION

A new version of the CLiDE software, CLiDE Pro, has been developed. After preprocessing a chemical image to identify connected components and graphical regions, a number of well-separated modules are involved in the reconstruction of chemical structures. At several stages of the reconstruction process, special rules are applied based on shape and contextual analysis of the connected components as well as chemical knowledge to detect and interpret difficult drawing features such as crossing bonds and resolve ambiguous situations.

The results reported in this paper show a high rate of accuracy in the reconstruction of molecules. The fact that high success rates are reached on each of the types of drawing features found in the test images indicates a consistent performance of the individual modules designed to deal with these features. Tests also show that the performance is greatly affected by the kind of noise which causes components which should be separate to be merged

together, e.g. a dash of a dashed bond touching a character of an atom label. The interpretation of merged connected components is very difficult requiring the detection of the attachment points where the connected components are joined together and the determination of each connected component's role in the structure such as a bond line or a character.

Similar tests were carried out on the original CLiDE software over about half of the images included in the test set. Results show that significant improvement has been achieved in CLiDE Pro regarding the following areas: (a) reconstruction of atom labels containing more than one character, (b) interpretation of chlorine and iodine atoms and formal atom charges, (c) interpretation of short bond lines as bonds (rather than characters of atom labels), (d) reconstruction of wavy bonds, (e) identification of dashes of dashed bonds, and (f) identification of aromatic ring circles.

The interpretation of generic structures is a very useful but complex task involving the following steps: (1) interpretation of the structure diagrams including super atoms containing R-groups, (2) identification of generic text blocks, (3) interpretation of the generic text blocks to identify the R-groups and their substituents, and (4) finding a generic match between the molecules and the generic text blocks. Initial tests show that CLiDE Pro works well on examples for which its implementation is designed. However, the steps are performed sequentially in CLiDE Pro. This means that if any of the steps fails or produces errors, the final reconstruction of a generic structure is very likely to fail. For example, tests show that generic text blocks lying close to structure diagrams are included in the graphical blocks surrounding the structure diagrams during document image segmentation. Therefore they are not detected as text blocks containing R-group information and subsequently not subjected to the generic text interpreter. Work on the segmentation of graphical blocks to identify text blocks that are not used in the compilation of connection tables is currently under way.

At present, CLiDE Pro is able to interpret R-group assignments expressed in linear text format including special symbols separating R-groups and substitution values. The implementation of a universal generic text interpreter is required in order to interpret R-group assignments defined in various formats such as tables of partial structure diagrams representing substituents to R-groups. The recognition of further alternatives of variability, namely position variation and frequency variation, often used in Markush structures is also one of our goals for the future.

The complete set of test images used for validation studies is available in the Supporting Information. We strongly recommend that developers of other OCSR systems should validate them using the same test set in order to provide an honest and clear comparison of the merits of their systems. We suggest that this test set could be the starting point for a community-based effort to establish a benchmarking test set which would include different categories of images each of which dealt with specific problem types.

Supporting Information Available: Test images used for validation studies. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) McDaniel, J. R.; Balmuth, J. R. Kekulé: OCR-Optical Chemical (Structure) Recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (4), 373–378.

(2) Borman, S. New Computer Program Reads, Interprets Chemical Structures. *Chem. Eng. News* **1992**, *70* (4), 17–19.

(3) Contreras, M. L.; Allendes, C.; Alvarez, L. T.; Rozas, R. Computational Perception and Recognition of Digitized Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (3), 302–307.

(4) Casey, R.; Boyer, S.; Healey, P.; Miller, A.; Oudot, B.; Zilles, K. Optical Recognition of Chemical Graphics. In *Proc. 2nd Int. Conf. on Doc. Anal. and Recogn. (ICDAR'93)*, Tsukuba Science City, Japan, Oct 20−22, 1993; pp 627−631.

(5) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A. G.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, *33* (3), 338–344.

(6) Zimmermann, M.; Le, T. B. T.; Hofmann-Apitius, M. Combating Illiteracy in Chemistry: Towards Computer-based Chemical Structure Reconstruction. *ERCIM News* **2005**, *60*, 40–41.

(7) Algorri, M. E.; Zimmermann, M.; Hofmann-Apitius, M. Automatic Recognition of Chemical Images. In *ENC 2007: Eighth Mexican International Conference on Current Trends in Computer Science*, Morelia, Mexico, Sep 24−28, 2007; IEEE Computer Soc.: Los Alamitos, California, U.S.A., 2007; pp 41−46.

(8) Algorri, M. E.; Zimmermann, M.; Hofmann-Apitius, M. Reconstruction of Chemical Molecules from Images. In *Proc. of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, Aug 22−26, 2007; IEEE: New York, 2007; pp 4609−4612.

(9) Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; Saitou, K. Automated Extraction of Chemical Structure Information from Digital Raster Images. *Chem. Cent. J.* **2009**, *3*, 4.

(10) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740−743.

(11) Tetsuo, A.; Danny, Z. C.; Naoki, K.; Takeshi, T. Polynomial-time solutions to image segmentation. In *Proc. of the seventh annual ACM-SIAM symposium on Discrete algorithms*; Society for Industrial and Applied Mathematics: Atlanta, Georgia, United States, 1996.

(12) Simon, A. Image Analysis and Content Retrieval of Printed Chemistry Documents, Ph.D. Dissertation, School of Chemistry, University of Leeds, 1996.

(13) Aho, A. V.; Hopcroft, J. E.; Ullman, J. D. *Data Structures and Algorithms*; Addison-Wesley Publishing Company: Reading, MA, 1983.

(14) Venczel, T. Recognition of Primitives, Ph.D. Dissertation, School of Chemistry, University of Leeds, 1993.

(15) Sklansky, J.; Gonzalez, V. Fast Polygonal Approximation of Digitized Curves. *Pattern Recognit.* **1980**, *12*, 327–331.

(16) Illingworth, J.; Kitter, J. The Adaptive Hough Transform. *Comput. Vision, Graphics Image Process* **1988**, *44* (2), 87–116.

(17) Duda, R. O.; Hart, P. E. Use of the Hough Transform to Detect Lines and Curves in Pictures. *Graphics Image Process.* **1972**, *1* (2), 409–418.

(18) Kam, F. Automated Extraction of Chemical Information from the Chemical Literature, Ph.D. Dissertation, School of Chemistry, University of Leeds, 1994.

(19) Kam, F.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical Literature Data Extraction - Bond Crossing in Single and Multiple Structures. *Recent Advances in Chemical Information II*, Annecy, France, Oct 19−21, 1992; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, United Kingdom, 1993; pp 113−126.

CI800449T