Improvements in Optical Structure Recognition Application

Igor V. Filippov
Chemical Biology Laboratory
SAIC-Frederick, Inc.,
NCI-Frederick
Frederick, Maryland 21702
igorf@helix.nih.gov

Marc C. Nicklaus
Chemical Biology Laboratory
NCI, NIH, DHHS,
NCI-Frederick
Frederick, Maryland 21702
mn1@helix.nih.gov

John Kinney
DuPont Stine Haskell
Research Laboratories
1090 Elkton Road, Newark,
Delaware 19711
John.B.Kinney@USA.dupont.com

ABSTRACT

We present recent improvements of the Optical Structure Recognition Application (OSRA), an open source utility to convert images of chemical structures to connection table type description in an established computerized molecular format. There exists a large body of chemical information which has remained largely inaccessible to machine data mining techniques so far. One of the most common ways of describing a chemical structure in a journal publication or a patent document is by drawing a two-dimensional structure diagram which represents atoms and bonds of the molecule in a human-recognizable form. While easily interpreted by a human expert, such drawings are by themselves unsuitable for use in a computer database for applications such as virtual screening and computer aided drug development. OSRA allows recognition and conversion of such drawings into computer formats widely used by the chemoinformatics community. This paper describes recent progress we have achieved for OSRA in terms of faster processing times and more accurate recognition rates.

Categories and Subject Descriptors

I.7.5 [**Computing Methodologies**]: Document and Text Processing—*Document Capture*; J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Chemistry*

General Terms

Design Performance Measurement

Keywords

Chemical structure recognition, Image extraction, Graphics recognition, Line drawings, Technical drawings

1. INTRODUCTION

Optical Structure Recognition Application is an open source utility which automatically detects, extracts and converts images of chemical structures, such as published in journal articles and patents, into machine readable formats such as SMILES (Simplified Molecular Input Line Entry Specification) or SDF (Structure Data File) formats.

The general work-flow of OSRA has been presented earlier [9] [10]. A page image is first segmented - all pairwise Chebyshev distances between connected components are calculated (unless they were estimated to exceed 50 pixels) and a threshold value is computed such that chemical diagrams and corresponding characters forming atomic labels are grouped together, while text and linear vertical or horizontal page separators are removed. This threshold distance is estimated based on emperical relationship between the size ratio of the connected components vs. their distance from each other. For more detailed description see [10]. Thinning and anisotropic smoothing are applied when necessary [8, 1] - that is, images processed at a resolution of 300 dpi or higher are subject to the thinning procedure, the anisotropic smoothing is applied when the noise factor is found to exceed an emperically found threshold. A noise factor is defined here as a ratio of the number of linear pixel segments (vertical or horizontal) with a length of 2 pixels to the number of line segments with a length of 3 pixels. The molecular diagrams are then vectorized using the Potrace library [4]. Connected components of certain aspect values, fill ratios and sizes are processed through two open source OCR engines, GOCR [3] and OCRAD [2], to determine atomic labels. Such processing is done at several different scales (three in earlier versions, four in recent releases) to allow for different possible scanning resolutions. A connection table is constructed to represent a molecular object by the OpenBabel chemoinformatics library. The best candidate structure is selected by using an empirically found confidence estimate function which gives a numerical score based on a number of simple molecular descriptors - common chemical element counts, number of rings, number of aromatic rings, fragment counts etc. This numerical score is in general higher for more chemically “sound” candidate structures. More detailed description has been published in [9].

Multi-page PDF and PostScript documents can also be processed, the only difference being that the density (rendering resolution) is preset to a certain value for efficiency reasons instead of attempting several different scales.

2. RECOGNITION ENHANCEMENTS

DAS '10, June 9-11, 2010, Boston, MA, USA
Copyright 2010 by the author(s)/owner(s) 978-1-60558-773-8/10/06

A number of enhancements were added to the recognition engine after release 1.2.2, which resulted in better recognition, and, in some cases, in making it possible to process documents which were not parsable at all before. Previous versions of OSRA by default attempted to process an image at three different scales (for non-PDF or PostScript files): 72, 150 and 300 dpi. Therefore documents scanned at a resolution higher than 300 dpi were either impossible to process or produced poor results with the default settings. This was rectified starting with release 1.3.0 — a fourth scale has been added which is dynamically adjusted according to the following algorithm: A histogram of run lengths of vertical and horizontal line segments is computed. If the position of the maximum is located above 6 pixels, the resolution is estimated as $300\frac{P_{max}}{4}$, otherwise the fourth resolution is taken to be 500 dpi. This allows processing of such images as, for example, the recent CLiDE Pro validation set, which seems to have been scanned at a resolution greater than 300 dpi, without having to worry about the specific settings that were used for scanning these images.

Another important improvement is the introduction of an option to process PDF and PostScript documents at other than the default resolution of 150 dpi, and the ability to rotate the page before processing. In earlier versions, an attempt to process a PDF document at 300 dpi density led to a runaway memory leak or even a program crash. These issues were identified by using the Valgrind memory debugger and were fixed starting with release 1.3.1. This allows better processing of documents such as USPTO PDF files, which sometimes contain structures rotated by 90 degrees and a typography style with overly thin lines when used at densities lower than 300 dpi.

In release 1.3.3, the page segmentation algorithm was updated to identify and remove tables and table-like graphical elements which sometimes are used in the documents to organize molecular images visually. Such table-like constructs often confused the segmentation procedure in the previous releases to the point were the processing time was greatly increased or no chemical structure could be identified at all. A table is identified as a connected component which is greater than 300 pixels in size, has an aspect ratio between 0.1 and 10, and at least 100 pixels of which are lying on the surrounding rectangle. An example of such table removal is shown in Figure 1.

3. PERFORMANCE IMPROVEMENTS

Significant speed improvements have been introduced in version 1.3.1 and especially 1.3.2 of the software. OSRA was profiled using the Gprof open source software and the code improved to remove bottlenecks. Further improvement was achieved by replacing the ImageMagick library with a more optimized GraphicsMagick.

The results are presented in Table 1. CLiDE validation sets are available from the SimBioSys Inc. website [5]. The small set consists of 14 images scanned at a resolution of 300 dpi. "USPTO CWU" is a collection of 5735 complex work unit files from the US Patent Office, consisting of image files and associated MOL files with molecular structures. Each image file was selected to contain 1 chemical structure. This set was prepared in collaboration with Dr. Steve Boyer and is available for download from OSRA website [6]. "WIPO PDF" is the patent WO2009009041A2 presented as a 192 page long PDF document and processed with the default settings, while the "USPTO PDF" file US20090176770(A1) has been rotated by 90 degrees and processed at a resolution of 300 dpi (options -R 90 -r 300). These files can be found on WIPO and USPTO web portals respectively.

In all cases the output was saved as SD files and the environment variable OMP_NUM_THREADS was set to 1. A desktop PC with an i7 950 CPU (3.07 Ghz) running Fedora 12 Linux x86_64 was used for benchmarking. All times are reported in seconds.

Starting with version 1.3.5 support for multi-threaded processing of PDF files has been added (currently only for Linux version). Each page is processed in its own thread which leads to a significant performance improvement on a modern multi-core system. The speed-up results are presented in Table 2.

Finally, the recognition rates for two sets of images where the ground truth was available are presented in Table 3.

4. CONCLUSIONS

We have presented our latest release of the Optical Structure Recognition Application. A strong effort has been made to make the code more robust, efficient, and more widely applicable in real world scenarios. A significant speed up (up to 3 times) has been achieved, while also increasing the scope of usability to higher resolution images, images with tables, and improving the recognition of PDF documents. OSRA is currently being used by industry and government organizations to process patent and scientific publication documents. It is available for free as source code as well as Microsoft Windows and Mac OS X installable executable files [6, 7]. Plugins are available for integration with popular molecular editor programs, such as Symyx Draw, ChemBioDraw, and BKChem.

Funding Disclaimer:

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

5. REFERENCES

[1] *GREYCstoration.* http://www.greyc.ensicaen.fr/ dtschump/greycstoration/

[2] *Ocrad - GNU Project - Free Software Foundation (FSF).* http://www.gnu.org/software/ocrad/ocrad.html

[3] *Optical Character Recognition (GOCR).* http://sourceforge.net/projects/jocr/

[4] *Peter Selinger: Potrace.* http://potrace.sourceforge.net/

[5] *SimBioSys Inc. CLiDE Validation.* http://www.simbiosys.ca/clide/validation.html

[6] *OSRA: Optical Structure Recognition.* http://cactus.nci.nih.gov/osra/

[7] *SourceForge.net: OSRA.* http://sourceforge.net/projects/osra/

[8] J. M. Cychosz. Efficient binary image thinning using neighborhood maps. In *Graphics gems IV*, pages 465–473. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[9] I. V. Filippov and M. C. Nicklaus. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *Journal of Chemical Information and Modeling*, 49(3):740–743, MAR 2009.

[10] I. V. Filippov and M. C. Nicklaus Extracting chemical structure information: Optical structure recognition application. In *Proceedings of the Eight IAPR International Workshop on Graphics Recognition*, pages 133–142, 2009.

Table 1: Performance benchmarks (time in seconds)

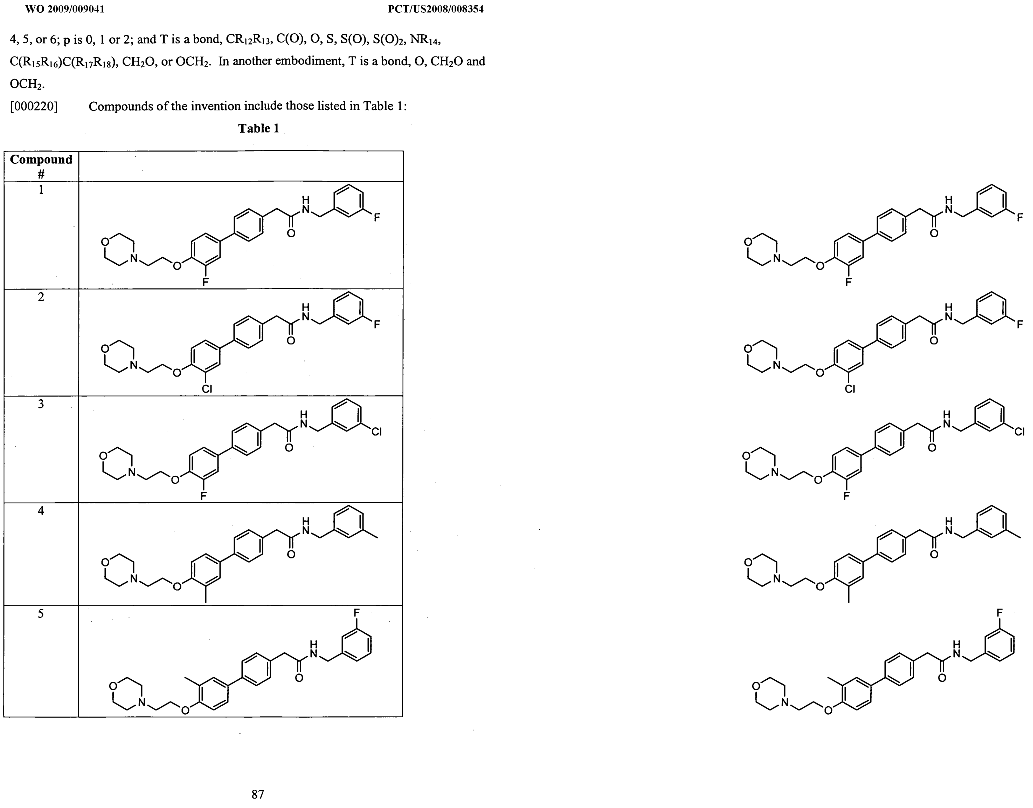
Test set	Number of pages	Total time v1.2.2	Time per page v1.2.2	Total time v1.3.6	Time per page v1.3.6
CLiDE small set	14	266	19	102	7.2
USPTO CWU	5735	23081	4.0	11222	1.9
WIPO PDF	192	—	—	486	2.5
USPTO PDF	314	—	—	2993	9.5

Table 2: Multi-threaded scaling for WIPO and USPTO PDF documents

Number of cores	2	3	4
WO2009009041A2	1.65	2.25	2.53
US20090176770(A1)	1.95	2.64	3.61

Table 3: Recognition Rates

Set/Version	Structures	v. 1.2.2	v. 1.3.6
CLiDE small set	46	26 (56%)	27 (58%)
USPTO CWU	5735	3984 (69%)	4341 (75%)



WO 2009/009041 PCT/US2008/008354

4, 5, or 6; p is 0, 1 or 2; and T is a bond, $CR_{12}R_{13}$, C(O), O, S, S(O), $S(O)_2$, NR_{14}, $C(R_{15}R_{16})C(R_{17}R_{18})$, CH_2O, or OCH_2. In another embodiment, T is a bond, O, CH_2O and OCH_2.

[000220] Compounds of the invention include those listed in Table 1:

Table 1

Compound #	
1	
2	
3	
4	
5	

87

(a) Original page

(b) Image after segmentation

Figure 1: A sample page from a WIPO document