Page Segmentation for chemistry Journals

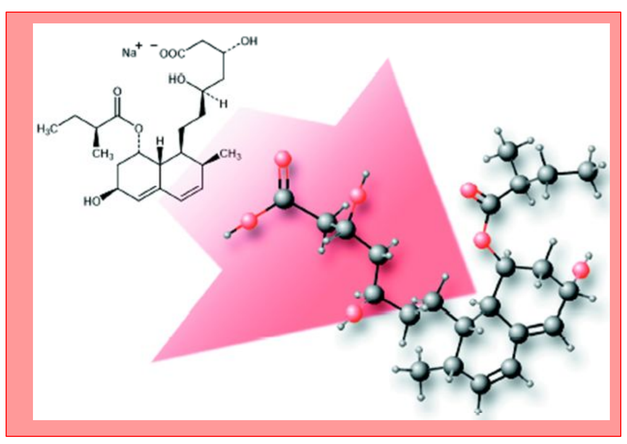
Marc Zimmermann[+], Nils Frings*, Sebastian Ginzel[#], Carina Haupt[#] and Christoph Friedrich[+]

[+] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics; Germany
* University of Applied Sciences Koblenz, RheinAhrCampus Remagen; Germany
[#] University of Applied Sciences Bonn-Rhein-Sieg; Germany

Keywords: Information Extraction, Page Segmentation, Image Classification, Machine Learning

September 2008

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen



Overview:

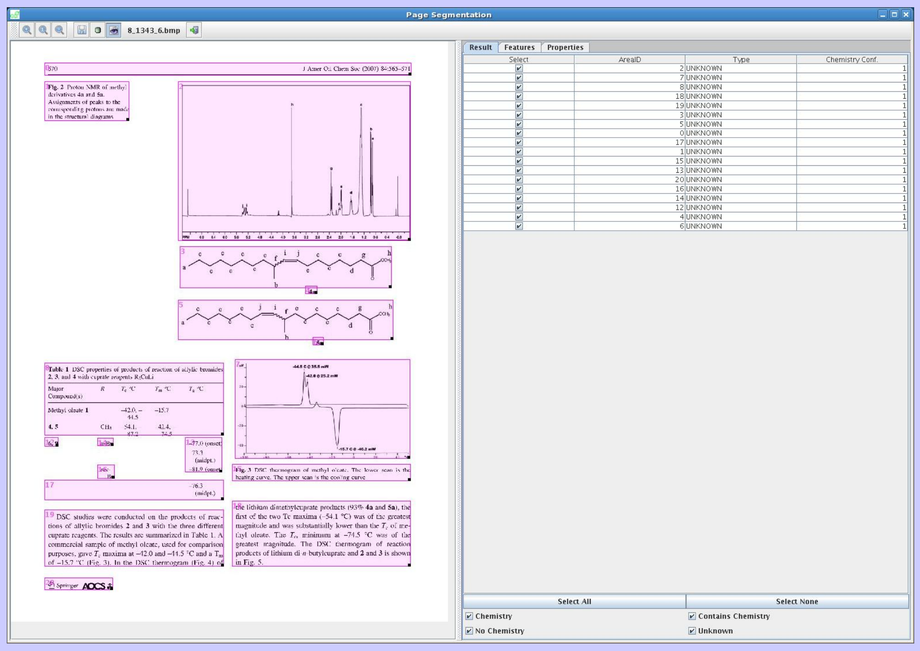
Scientific literature contains a lot of information which is found not only in text but also in depictions, tables and diagrams. Due to the fact the layout of scientific articles is quite complicated (like this poster) and differs from normal text sources.
Nowadays chemistry journals are already digitally available (e.g. as PDF). However for extracting chemical related information like structural formulas from these sources it is necessary to identify and separate chemical objects and their context from the rest of the document. In the second step the separated chemistry objects are processed by name-to-structure or image-to-structure tools like *chemoCR*[TM].

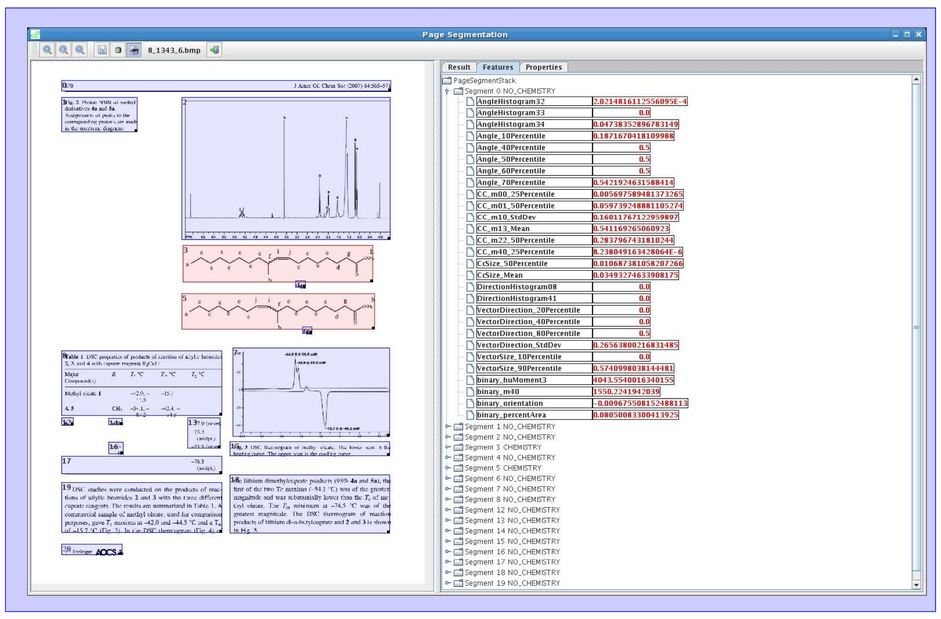
Challenge 1: grouping of semantic entities in full page scans

The first step is to find all coherent regions and mark them with boxes. It is possible that several chemical structures appear in such a region. However, it is not allowed to have text, figures and structures mixed in one region. Coherent regions should consist of one chemical structure, one diagram, one text paragraph, a caption or section header. Columns should be separated.

Methods:
Input to the segmentation module are grayscale and binary full pages scans from PDF documents. An algorithm coined *region growing* has been developed. It uses *dilation* and *connected component* algorithms. The dilation filter leads to a fusion of small neighboring connected components. A connected component is a collection of foreground pixels which are next to each other. Depending on the resolution of the image and the font types used, different kernel sizes and iteration numbers are automatically selected.

Result:
The screenshot to the right displays the regions of interest (ROI) in pink boxes. Each diagram, each chemical structure, each text paragraph form its own ROI. The table has been dissected into multiple parts.





Challenge 2: classification of semantic entities in full page scans

The second step is to label the correct regions in the image as containing chemical structures (CHEMISTRY) or not containing chemical structures (NO_CHEMISTRY).

Methods:
We assembled a training set of about 5000 images containing a chemical structure and about 5000 diagrams, figures, tables and text images without chemistry. All of them have been binarized and manually annotated. 6 different classification algorithms have been trained on the training set. These are: *naïve bayes*, *kNN*, *SVM*, *C4.5 tree*, *random forest* and *neuronal net*. The models have been optimized by applying *feature normalization* and *selection*, *bagging* and *parameter sampling*. For each image a feature vector has been computed containing binary features (moments), angle distributions, line and orientation features and connected component features.

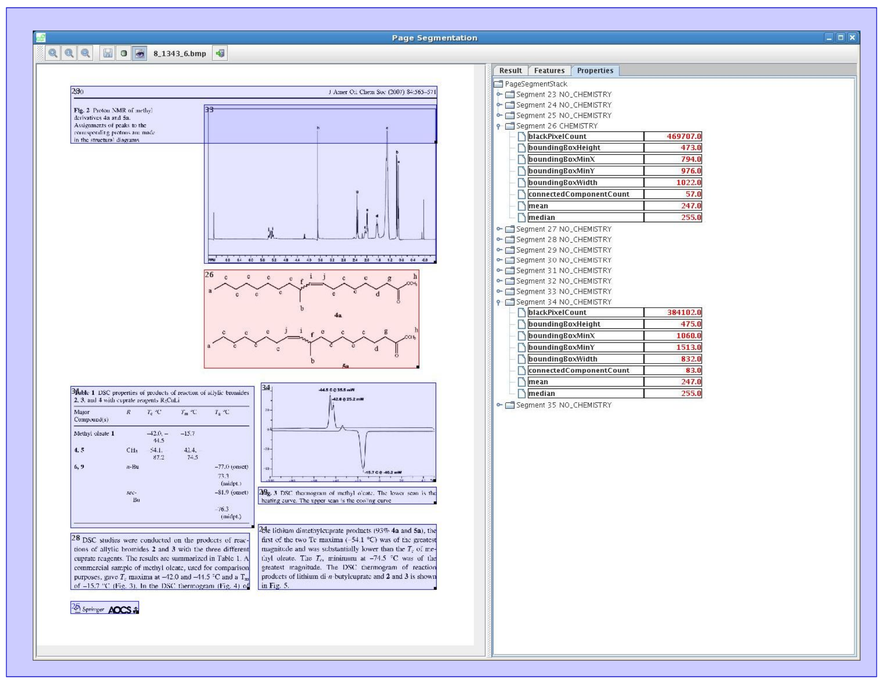
Result:
The accuracy of the different algorithms on the training set is between 92% (naïve bayes) and 97% (random forest).

The screenshot above shows the result of the classification of the ROIs. The two chemical structures have been correctly classified. The chosen features can be seen on the right panel.

Challenge 3: clustering of semantic entities and their context

A lot of chemical depictions contain more than one chemical structure (e.g. reaction schemes and tables). These should be passed to *chemoCR*[TM] as one input image. Therefore we cluster chemical structures with neighboring objects into larger boxes.

Methods:
The clustering is rule based. It merges chemical objects if the neighboring object is also chemical or if the neighboring object is small enough. There is only one parameter the *mergeSizeLimit*. Small non chemistry objects which have to been merged are normally reaction arrows or annotations of chemical structures. The final ROI is sending its connected components to *chemoCR*[TM] for further reconstruction. Overlapping parts from other ROIs are ignored (like in the screenshot to the right).



Proof of Concept:

For an evaluation 440 full page scans have been selected from 15 different journals on chemistry published on SpringerLink. In total 223 chemical ROIs have been manually cut out and used as gold standard. For each ROI the bounding box, number of foreground pixels, connected components, mean and median color has been computed and compared to the gold standard. The computed ROI properties can be seen in the right panel in the upper screenshot.

Non chemistry ROI as chemical ROI (false positive)	Good match of chemical ROI (>90% fit)	Partial match of chemical ROI (60% - 90% fit)	Chemical ROI not identified (false negative)
11	123 / 223	97 / 223	3 / 223

References:
[1] Frings, N.; "*Vergleich von Bildklassifikationsalgorithmen für chemische Strukturformeln*"; Bachelor Thesis, Biomathematics, University of Applied Sciences Koblenz, RheinAhrCampus Remagen; Germany; 2008.
[2] Ginzel, S. and Haupt C.; "*Identifying chemical structures in publications*"; Computer Vision Project Report, University of Applied Sciences Bonn-Rhein-Sieg; Germany; 2008.

Contact
Fraunhofer Institute SCAI.Bio
E-Mail: marc.zimmermann@scai.fraunhofer.de
URL: http://www.scai.fraunhofer.de/chemocr.html