Recognising Chemical Formulas from Molecule Depictions

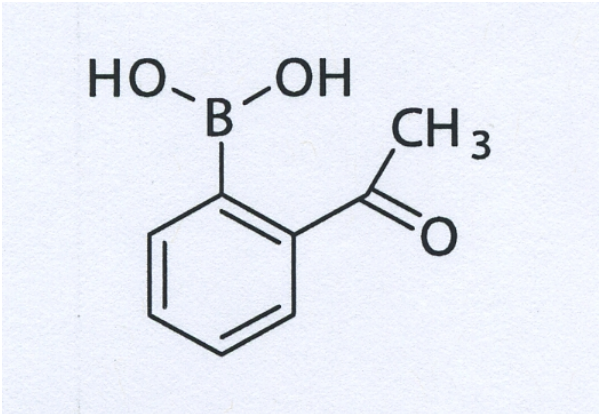
Noureddin Sadawi

School of Computer Science
University of Birmingham

GREC 2009, La Rochelle, 22nd July 2009

Introduction

- Working on Document Analysis (Chemical Documents).
- Putting together a set of tools (NMR Parser, Molecular Diagram Recogniser)
- Not too many existing tools, few open source (e.g. OSRA).
- Some tools depend on other tools, not easy to integrate.



OB(O)C(C=C1)=C(C=C1)C(C)=O

Molecule Recognition Procedure

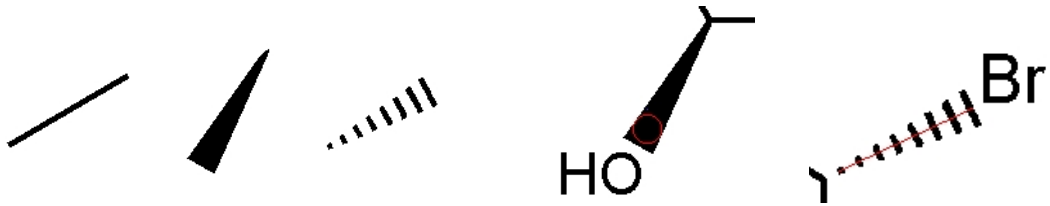
- Thresholding (Otsu's method) & Connect. Comp. Labelling

Molecule Recognition Procedure

- Thresholding (Otsu's method) & Connect. Comp. Labelling
- OCR via Zoning
  Indentify character candidates & extract numerical features via a 3×3 grid.

Molecule Recognition Procedure

- Thresholding (Otsu's method) & Connect. Comp. Labelling
- OCR via Zoning
- 3D bond extraction (solid triangles & dashed bonds)

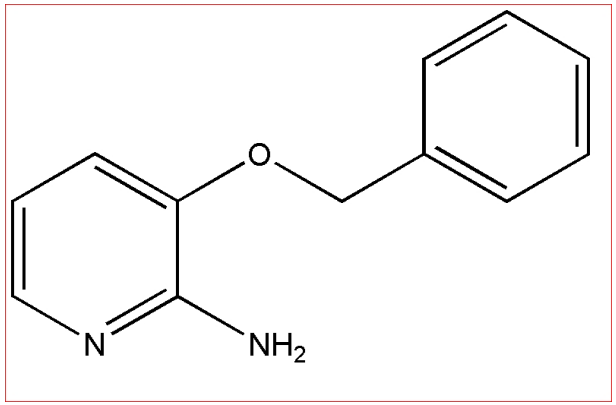


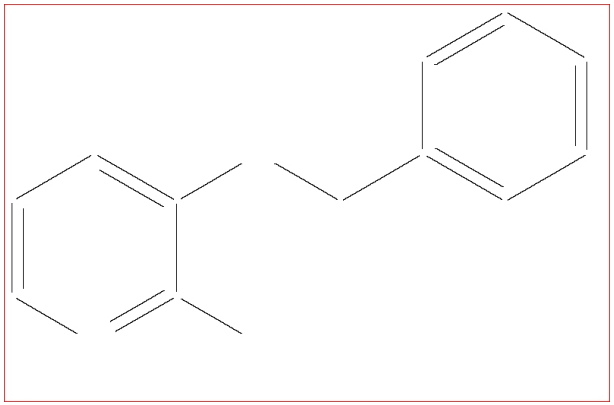
Molecule Recognition Procedure

- Thresholding (Otsu's method) & Connect. Comp. Labelling
- OCR via Zoning
- 3D bond extraction (solid triangles & dashed bonds)
- Characters removal & Thinning (Hilditch's algorithm)

Molecule Recognition Procedure

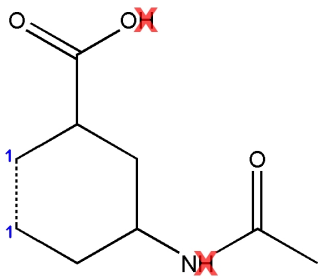
- Thresholding (Otsu's method) & Connect. Comp. Labelling
- OCR via Zoning
- 3D bond extraction (solid triangles & dashed bonds)
- Characters removal & Thinning (Hilditch's algorithm)
- Disconnect line segments and construct labelled graph

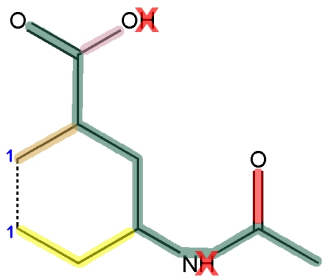




Molecule Recognition Procedure

- Thresholding (Otsu's method) & Connect. Comp. Labelling
- OCR via Zoning
- 3D bond extraction (solid triangles & dashed bonds)
- Characters removal & Thinning (Hilditch's algorithm)
- Disconnect line segments and construct labelled graph
- Graph Traversal to extract SMILES notation





O=C(O)C(C1)CC(CC1)NC(=O)C

Results and Discussion

- 60 different images images created with the chemical software package ChemDraw.
  - No noise.
  - 59 recognised correctly.
  - 1 recognised incorrectly (Due to drawback in thinning algorithm).

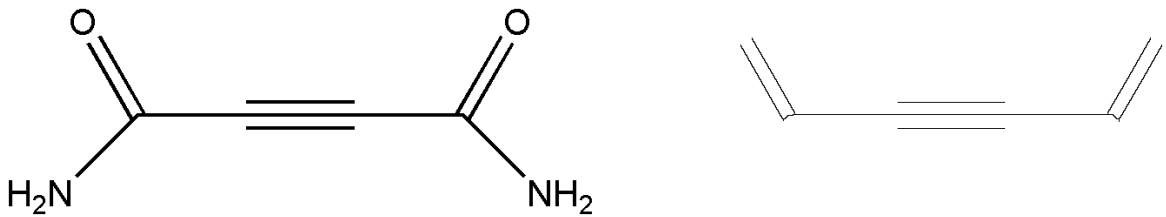
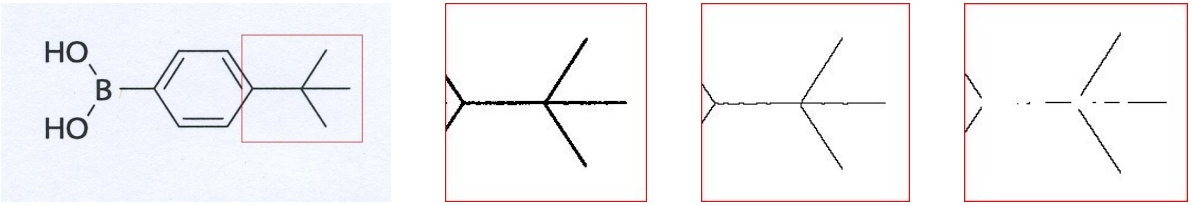


Figure: Problem with Hilditch's thinning algorithm

Results and Discussion

- 50 scanned images from different sources (different catalogues).
- 23 were recognised correctly.
- 24 recognised incorrectly (lines disconnected at undesired locations).



- 2 recognised incorrectly (Iodine atom classification error).
- 1 recognised incorrectly (diagonal line completely erased after thinning).

Conclusion & Further work

Work done so far:

- Used Hough transform and line slopes to extract precise data but results were inaccurate.
- Used simple approach and had encouraging results.
- Built graph to represent molecule structures.
- Extracted meaningful representation.

Future work:

- Use another thinning algorithm & improve accuracy of line junction/end-point detection.
- Handle more complicated images.
- Generate LaTeXsource as output.
- Handle whole document and automate clipping.