印刷された化学構造式の認識理解システム

伊藤尚樹
中山　尭

Image Understanding System for Printed Chemical Structures

Naoki Itoh
Takashi Nakayama

This paper describes an image understanding system which reads printed chemical structures via an image scanner, recognizes them as chemical structures and gives the result in the form of connection tables. The system accepts chemical structure diagrams including characters and stereochemical symbols depicted in ordinary textbooks and articles. The recognition and understanding process of image data comprises the following three major steps: at first, the basic recognition of building blocks of input data as figure primitives, then the graph generation from the result of the previous process, and the interpretation and understanding of the graph as a chemical graph (a chemical structure). In other words, the whole process consists of hierarchically distributed functions, where the bottom layer is the first one, and the top the last. The process normally progresses from the bottom to the top. However, the system works cooperatively between the layers when some layer suspends judgements or detects the possibility of misjudgements. For instance, an assumed character area could be ambiguous when the character is overlapped with other figure elements; crossing lines may be interpreted as a cross of two single lines, or four lines connected to a single node. These undetermined situations can be resolved with cooperative processing of two or more layer functions using domain knowledge concerning chemical structures. Experiments have proved that the system is powerful enough to give sufficient precision in terms of center-line extraction of complex chemical structures which a commercial product cannot recognize precisely.

1　はじめに

文書画像を自動的に認識理解するシステムの研究が種々の具体例を対象として行なわれるようになっている。[1][2] ここでは印刷された化学構造式（立体構造式を含む）を対象として，それを化学構造として解釈するシステムに関する研究について報告する．同様の研究はすでにいくつか発表されており，市販ソフトも供給されている．[3] しかし，性能的には改善すべき部分がまだ多く残されていると思われる．たとえば，線分（結合）が入り組んで図形的に複雑になっているような場合，文字（原子記号）と線分が接触したり交わっている場合，あるいは複数線分の作る角のつぶれや 短い線分と点線破線の単位との区別の認識などについては明らかに限界がある．[4] これらの問題は他の文書画像の認識理解においても共通した課題であり，その解決のために様々のアプローチが試みられている．本研究では対象の認識と理解のプロセスがいくつかの機能の協調と統合によって実現されているという仮定に立って，階層的な分散協調システムとしてこれを実装することを試みた．ここでは主としてその中の線図形の抽出と認識について述べる．

2　化学構造式認識

化学構造式を認識する上で現在の主な問題点として次の２つがある．1. 複雑な線図形の認識：結合が立体的にこみいった構造を平面上に描こうとすると，結合を示す線分のために十分な長さをとることができなくなる．文字と結合を分離する情報として単純にはそれら線分のサイズを利用するので，このような場合には文字認識にも不利な状況となる．また，立体構造を表すための線分の交差や，さらにその交差部分に描かれた点線や破線と短い結合との誤認識が正しい認識を困難としている．2. 文字と結合の重なり：本来は文字と結合は接触しないように描かれると期待したいところであるが，現実には重なっている場合

も多い．文字領域の切り出しを行なうのに文字領域のサイズ情報のみに着目すると，これらの分離・抽出が困難となる．

これらの問題を図形（画像情報としての化学構造）の局所的な特徴，すなわち着目している図形部分にのみ基づいて解決するのは困難であり，問題部分の周囲の線図形構造まで着目範囲を広げて，化学構造式に関する専門知識を適切に用いる必要が生じる．これらのことを考慮し，本研究では線図形特徴から化学構造式を認識する過程に段階を設け，下層に図形のプリミティブな要素の認識や生成機能を持たせ，上層に下層の情報に基づいて全体の画像情報を化学構造式として解釈させる機能を持たせるという機能分散を行なうというような，ボトムアップとトップダウンの協調的な認識理解方式を提案する．また，この方式を実現するために線図形の情報を抽出する線図形構造解析と文字認識の機能についてはすでに開発済みである．

本システムはイメージデータから線図形要素を抽出する線図形構造解析モジュール，抽出された線図形要素を段階的に理解し，化学構造式として認識する化学構造式認識モジュール，またこれら2つのモジュールから依頼を受けて，想定された文字領域において文字認識を行なう文字認識モジュールの3つから構成される．以下に階層化された認識方式の概要と線図形構造解析および文字認識モジュールについて述べる．

3　階層化された化学構造式認識

一般に，構造を持つ情報はより粒度の細かい部品によって構成されている．またその部品にも構造がありさらに粒度の細かい部品によって構成されることが多い．このようにして構造を持つ情報を細分化していくと，それ以上意味を有する部品に分割できなくなるプリミティブな部品に到達する．逆に，このプリミティブな部品を処理の出発点とすると，ルールまたは知識を用いて部品から構造情報を生成していくことができる．構造情報の解析の過程，あるいは部品から構造情報を生成する過程を機能階層 (layer) として定義し，この解析と生成の2つの機能を実現する．各層には下層の部品から自層の構造情報を生成するルールあるいは知識と生成手段を，また上層からの要求に答えるための手段を持たせることにより独立性を保持すると同時に，下層との間で協調することにより，部品からの構造情報の生成という形でボトムアップに認識過程を進めて行く．下層との間の協調とは，自層の構造情報の生成段階においてすでに部分的に生成された自層の構造情報と下層の部品との間に矛盾が生じた場合に，その矛盾を解消するような構造の生成を下層に要求し，その妥当性と変更による影響の報告を受けて自層の構造情報を決定することを表す．また認識の柔軟性を考慮し，自層での構造情報の生成が一意に決定できない場合は上層に決定を委ねる．図1に本システムの構成を示す．

1. **図形層**　構造解析によって得られた図形要素に2次元空間情報を付加した情報を上の層に提供する．この層は次の3つの副層からなる．図形層の機能は線図形構造解析で得られた図形要素の間に幾何的意味を付加することである．たとえば，2つのベクトルの平行性，2つのベクトルの間の角度，ベクトルの両端や延長線上の最近傍に存在する図形要素の示唆などである．

ルート層　線図形構造解析で得られた全ての情報を提供する層である．すなわち，すべての情報を他の階層および副層で扱われるデータ構造（オブジェクト）に変換する．この際，連結情報のような既に保有している構造情報はそのまま保存される．

図形記述子層　ルート層の情報に基づいて広がりなどの2次元空間情報，たとえば，各図形要素の代表点間の位置関係などを付加する．位置関係は代表点間の近くにある他の代表点の図形要素とリンクを張ることにより実現する．またベクトルに関しては始点，中点，終点の3つを代表点とし，直接連結するベクトルはリンクの対象とはしない．

図形要素層　図形記述子層の情報に基づいて各ベクトルに対して図形としての意味を付加する．すなわち，直線，点線，破線，波線，クサビ型などの中のどれを構成しているかについて帰属を行なう．また，文字の平面上の配置を解釈して文字列の生成を行なう．この場合も，図形記述子層から渡される近傍情報を利用する．線図形構造解析で芯線化されなかった輪郭線についても，輪郭線の特徴に基づいてそれが線分に成り得るか否かの判断を行なう．特にクサビ型に関しては線分を形成する輪郭線は互いに平行ではないので，ペアベクトル探索の段階で線分候補から排除される場合がある．そこでこの層において輪郭線の特徴に着目してクサビ型としての抽出を行なう．

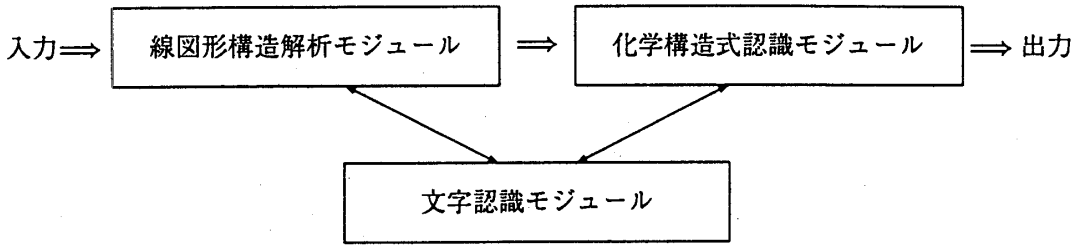

図 1: システムの構成

2. グラフ層 図形層から提供される情報に基づいてグラフを構築する．ノードになる要素とエッジになる要素の探索を行ない，それらの整合性をとりながらグラフを生成する．グラフ層は次の 2 つの副層を持つ．

グラフ要素層 グラフの要素であるノードとエッジを図形層から探索し定義する．非常に短い線分以外はエッジとする．化学構造式に固有の図形要素をグラフ要素として認識するために，それらのグラフ要素としての定義機能をこの層にもたせる．たとえば，クサビ型の線分や破線をエッジとし，明示的なノードが無いベクトルの端点や接合点にノードを生成することなどを行なう．また単独で存在する文字や文字列はノードの属性と判断しその場所にノードを生成する．文字と線分が接触する場合があるので，長い線分の端点に抽出済みの文字の大きさとほぼ同じ大きさの短い線分の集まりがあれば，それを分離し文字認識を行なう．文字認識の結果はノードの属性とする．

グラフ生成層 グラフ要素層のグラフ要素からグラフを生成する．極近傍に存在するノードの統合や消去，あるいは線分の統合なども行なう．具体的には，印刷が不明瞭で線分が掠れることがあるが，掠れた領域が狭い場合は端点同士が極近傍に現れかつ線分の角度が等しくなるので，ノードを消去し線分を統合することが可能となる．

3. 化学構造式層 グラフ層と図形要素層で生成された情報に基づいて化学構造式を認識する．グラフ層で生成されたグラフに化学構造式の情報を付加することにより化学構造式として認識する．この層は 2 つの副層からなる．

化学構造認識層 グラフ層のグラフを化学構造式として解釈する．すなわち，ドメイン知識を利用して化学構造として論理的に矛盾のないようにグラフを解釈する．したがって，ノードの属性は原子の属性とみなされる．ノードに属性が記述されていない場合は原子属性として炭素原子を付与する．

結合表生成層 認識された化学構造に基づいて結合表を生成し出力する．

4　線図形構造解析

線図形要素とは線画像の構成要素である線分に対して他の線分との接続関係や線分の太さなどの線分に関する特徴（属性）を付加したものである．画像内の文字も線図形要素として扱われる．原画像からの線図形要素の抽出は，1. エッジの抽出，2. エッジの折れ線近似，3. 芯線化，という手順で行なわれる．

エッジ抽出　イメージデータ上の線を表現する方法としてはフリーマンのチェインコードを用いる．[5] これは，線分をドットが連結したものであるとみなして，ドットの連結の仕方を隣接ドットに対する方向コード (チェイン) の列として表したものである．チェインの方向としては 8 近傍を考える (つまり，45°単位とする)．ここでは抽出したエッジをこの方法で表現する．エッジの抽出法は種々あるが，従来はエッジを抽出した後で連続要素の外側と内側のエッジの対応関係を決定するのに外接する矩形の大小関係と位置関係を利用している．しかし，この方法は本システムにおいては必ずしも十分ではない．すなわち，芯線化処理の段階で内外の対応つまり 1 つの連続要素を生成するエッジのグループ分けが必要となるので，本システムにおいてはイメージデータ上の連続要素が

画面の端に接していないという条件の下で，エッジを抽出しながら内外の対応関係を付ける方法を開発した．以下にそのアルゴリズムを示す．

1. イメージデータの左上隅から走査を始めて，0から1になる点を探す．この点は上述の条件から必ず外側のエッジ上の1点となる．
2. この点から進行方向に対して左側を内側にしてエッジを追跡してチェインコードを生成する．その際，エッジの一部となったドットに対して，抽出済みのフラグとしてそのドットが0ドットに接している面の方向（左右上下）と，それが外側のエッジになったことを示す値を設定する．
3. 内側エッジの開始点を見つける．生成されたエッジ上をチェインコードに従って走査し，チェインコードが1, 2, 3以外のドットに対して次の処理を行なう．はじめに，エッジの連続要素領域内にいることを示すフラグFを立てる．8連結の0方向に走査しドットのフラグ状態を調べて行き，Fが立っていて次が0ドットでかつ一度も右側がエッジとして抽出されていなければ，そのドットは内側のエッジの一部であるとする．このドットから出発してステップ2と同様にエッジを追跡して内側のチェインコードを生成する．0方向への走査は外接する矩形内でのみ行なう．

エッジの折れ線近似　折れ線近似の方法も多数提案されているが，本システムでは追跡法と統合法の2つを組み合わせる方式とした．具体的には，追跡法の1つであるコーン交差法[6]によって誤差が小さくなるように細かく折れ線近似を行ない，その結果に対して統合法を適用するものである．統合法は折れ曲がり方が小さい連続する線分を1つの線分に統合するものである．折れ曲がりの角度については閾値の設定によって調整が可能である．ここで得られる各線分を輪郭線ベクトルと呼ぶ．コーン交差法に対して厳しい閾値を設定するとより細かい線分の特徴が抽出される．したがって，これら2つの方法の閾値を調整することによって原画像の特徴をより忠実に反映するような近似線分の柔軟な抽出が可能となる．

芯線化　芯線化とは，並行する2本の輪郭ベクトル(ペアベクトル)を探索し，その中間に芯線となるベクトルを生成し骨格を得る方法である．この処理は連続要素の輪郭ベクトル毎つまり外側エッジの輪郭ベクトルとそれに対応する内側エッジの輪郭ベクトルのペアに対して行なわれる．線図形構造解析では，線分の幅も抽出されるので芯線化と同時に芯線化した線分の幅を抽出する．

5　文字認識

本システムにおいて認識対象とする文字カテゴリーは英数字といくつかの記号であるので，漢字などの場合に比較すれば精度の要求はそれほど厳しくはない．一方，最終的な目標は化学構造式の認識であるから，文字認識の処理が遅いと他の処理を高速化しても全体の効率が低下する．そのため，文字認識法としては計算量の点で有利なパターンマッチングを用いることとする．

文字認識の手法は数多く提案されているが，ここではイメージデータ上のドットの並びをそのまま文字パターン特徴量としてパターンマッチングを行なう方法を採用する.[7] 文字認識は一般にイメージデータからの特徴量抽出と特徴量を用いたマッチングの2つのプロセスに大きく分けられる．ここでの特徴量抽出プロセスは次のようになる．イメージデータ上のドットの並びから次の4つの特徴を抽出する．1つ目は縦横ともに64ドットの正方形のドットパターン特徴量，2つ目はこれを16ドットに圧縮したメッシュパターン特徴量である．さらに認識速度を改善するために本システムにおいて大分類のための3つ目の特徴量を導入した．この特徴量は4ドットに圧縮したメッシュパターンであり，文字カテゴリーを16グループに分類する．これをインデックス特徴量と呼ぶ．また，各グループ内の類似文字を抽出しそれらの違いを強調するために4つめの特徴量として差分ドットパターン特徴量を抽出する．文字認識モジュールは，線図形構造解析および化学構造式認識の両モジュールによって指定されたイメージデータ上の長方形領域に文字が存在するものと仮定して文字認識を行なう．メッシュパターン特徴量を抽出する際，従来の方法では文字領域の内部を忠実に走査するのみで領域の外縁部分の情報を十分反映しない結果となっていたが，本システムでは領域外縁の周囲2ドット列に0要素を付加して文字領域を仮想的に拡大し，拡大された領域を走査することでより忠実な特徴量の抽出を可能とした．

6　実験と考察

実験に用いたマシンはFujitsu S-4/2である．プログラムはC言語を用いて実装した．プログラムの総

ステップ数は約 12000 行である．このプログラムによる線図形構造解析と文字認識のそれぞれについての実験結果を以下に示す．

線図形構造解析　入力イメージのデータサイズは横 912 ドット，縦 590 ドットである．図 2 に示すような交差や点線の立体表記などを含む複雑な化学構造式の場合,[8] 各処理にかかった時間は，エッジ抽出 4.42 秒，エッジ折れ線近似 0.67 秒，芯線化 5.53 秒，文字認識 13.70 秒，合計 24.32 秒である．文字認識は 51 文字領域に対して行なわれている．このように複雑なイメージデータの場合も芯線化は正確に行なわれていることが確認された．

実験から次のことが主張できる．連続要素のエッジは内側，外側の判断を含めて正確に抽出することができた．芯線化は折れ線近似の結果に応じて正確さが影響を受けるが，実験で用いたデータ程度ならば良好に動作することを確認した．点線や波線も特徴を良く表した形状で抽出されている．したがって，イメージデータから化学構造式の線図形を抽出することは大部分の印刷データについて精度良く可能であることが分かった．この方法は細線化処理を利用した線図形構造解析よりも処理時間も大きく短縮できることも確認している．エッジ抽出は，最も時間のかかる処理であるが，基本的に画像のサイズに依存するため根本的なアルゴリズムの見直しを行なわないと大幅な短縮は困難であろうと考えられる．

文字認識　辞書作成データを作成するためのサンプル文字パターンは，TEX のフォント 2 種類をプリンターに出力し，再びスキャナーで読み込んだものを利用した．字種は，英字大文字と小文字 52 種，数字と記号 36 種の合計 88 種を用いた．カテゴリー数は 2 種のフォントを別のカテゴリーとして使用したので，字種の 2 倍の 176 カテゴリーを登録した．1 カテゴリーを生成するのに 5 文字の領域の平均を用いたので，入力した全ての文字領域は 440 個となる．辞書のサイズはおよそ 8 メガバイトとなった．作成時間は 200 秒程度である．

文字認識実験は上記の辞書を用いて，辞書を作成したのと同じフォント 2 種と辞書作成に用いなかったフォントの 1 種類を対象とし各フォント毎に行なった．カテゴリーは辞書のカテゴリーと同じものを用いて，各カテゴリーにつき 5 文字認識させた．文字のサイズは辞書作成時に利用したものと同じものを用いた．これらのパターンは紙に印字したものをスキャナーから 400dpi で読み込みイメージデータを作成した．また，文字領域は正確に抽出できるようにした．各フォント毎の認識率，1 文字あたりの認識時間は，Roman(辞書フォント)100%,0.266 秒，Typewriter(辞書フォント)100%,0.268 秒，SansSerif85.2%,0.266 秒である．SansSerif フォントは第 5 位認識率で 98.9%，第 8 位認識率で 100%となっている．

文字認識モジュールに関しては，次のことが主張できる．認識率の結果から辞書を作成したフォントと同種の文字に対しては文字領域の切り出しが完全ならば全く問題なく認識できることが分かる．また．異なるフォントを用いた認識でも第 5 位の認識結果でほぼ正確に認識できるので，上層の処理で誤認識の可能性があるとして文字認識モジュールに再び差し戻すことによって訂正できると考えられる．これが化学構造の知識を利用した処理の一例である．

以上のことから文字処理および線図形解析については化学構造の認識理解に必要とする十分な精度が得られたということができる．線分が入り組んで複雑になる場合，線分が交差する場合，あるいは文字と線が接触する場合など，従来のシステムで十分に認識できていないケースについても十分に処理が可能な精度である．これらの処理の結果を用いてさらに化学構造式の全体を矛盾のないように解釈することになる．

7　おわりに

文字認識では，大分類に用いるインデックス特徴量とメッシュパターン特徴量抽出時の補正方式を提案し，実装の結果それらの有用性が確かめられた．線図形構造解析では，輪郭ベクトル生成法にコーン交差法と合成法の 2 つを組み合わせることによって，より精度の高い線図形構造情報の抽出が行なえることを確認した．これらの結果，本システムで要求される情報の精度を十分に達成することができた．

また，線図形構造情報から化学構造式を認識する方法として階層的に機能分散する方式を提案した．この方式の特徴は，下位層に図形のプリミティブな要素の認識や生成機能を持たせ，化学構造に関する専門知識を利用することによって上位層に全体イメージを化学構造式として解釈させる機能を持たせるという機能分散が，ボトムアップとトップダウンの協調的な認識理解方式として自然に組み合わされることにある．これはまた様々な図形要素や文字・記号などが混在するシーンを認識するという人間の視覚機

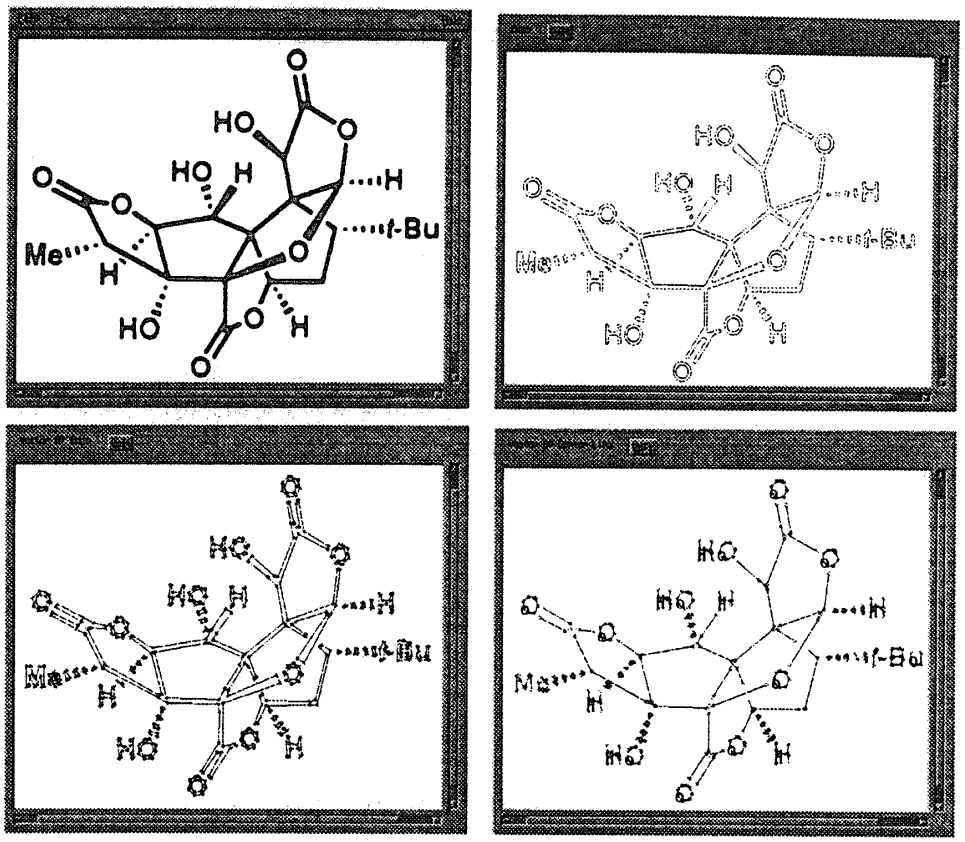

図 2: 複雑な化学構造式の線図形解析結果の例
左上: 入力イメージデータ, 右上: エッジ抽出
左下: エッジ折れ線近似, 右下: 線図形解析結果

能一般の自然なモデルであるとも考えられる. したがって, 印刷情報に限らず自然画像を含む様々の分野における画像情報の認識や理解に対して本研究の提案した方法が有効であると考えられる.

参考文献

[1] 角 保志 : 画像理解システムにおける分散的協調処理, 人工知能学会誌, 1994, 9, 5, 631-636.

[2] 黄瀬 浩一, 百田 賢一, 杉山 淳一, 馬場口 登, 手塚 慶一 : レイアウトとコンテンとの知識を用いた仮説駆動型文書画像処理, 情報処理学会論文誌, 1993, 3 4, 8, 1716-1729.

[3] Joe R.McDaniel and Jason R.Balmuth : Kekulé: OCR–Optical Chemical (Structure) Recognition, *J.Chem.Inf.Comput.Sci.*, 1992, 32, 373-378.

[4] 米田 幸夫 : 化学構造式考現学, *CICSJ Bulletin*, 1991, 9, 1, 68-74.

[5] H.Freeman : Boundary Encoding and Processing, in *Picture Processing and Psychopictorics*, B.S.Lipkin and A.Rosenfeld, Eds. Academic Press, New York, 1970, 241-266.

[6] C.M.Williams : An Efficient Algorithm for the Piece-wise Approximation of Plannar Curves, 1988, *CGIP*, Vol.2.

[7] 進藤 宣博, 阿曽 弘具, 木村 正行 : 低品質印字文字を高精度に識別する複合アルゴリズム, 情報処理学会論文誌, 1994, 35, 9, 1714-1721.

[8] E.J.Corey and Xue-Min Cheng: *The LOGIC of CHEMICAL SYNTHESIS*, WILEY-INTERSCIENCE, 1989.

神奈川大学理学部情報科学科 Department of Information Science, Faculty of Science, Kanagawa University