データ駆動型処理と概念駆動型処理の相互作用による文書画像レイアウト解析

石谷康人†

多様なレイアウト構造を持つ印刷文書を一括して読み取り，ディジタル化することを目的とした新しいレイアウト解析方式を提案する．本レイアウト解析方式は3つのデータ駆動型処理：領域統合，領域解析，領域認識と1つの概念駆動型処理：領域変更で構成されている．レイアウト解析システムでは処理モジュールが階層的に配置されており，隣接するモジュール間で相互作用を可能としている．前処理により抽出された文字成分がレイアウト解析システムに入力されると，データ駆動型処理の相互作用を経て，近接，類同，良い連続などの条件に基づいて文字成分が群化してクラスタを形成する．概念駆動型処理では，誤った群化を解消するための仮説をそれぞれのクラスタに適用する．そして，データ駆動型処理と概念駆動型処理の相互作用により仮説が検証され，システムは最終的に，テキストブロックと文字行と文字の階層関係で記述されるレイアウト要素の集合を文書画像のレイアウト構造として出力する．新聞，雑誌，科学技術文献，官報，表形式文書，書籍，名刺などを対象とした実験では，本方式が多様なレイアウト構造に適応していることと，複雑なレイアウト構造を高精度に処理できることを確認することができた．

Document Layout Analysis by Interaction between Data-driven Processing and Concept-driven Processing

Yasuto Ishitani†

A new method of document layout analysis is proposed for a document reader to be used for reading a wide variety of documents. The proposed system consists of three data-driven processes, i.e., region grouping, region analysis, and region recognition, and one concept-driven process, i.e., region modification. Each processing module can communicate with the adjacent modules, which means there is an interaction between modules. The interaction between data-driven processing modules organizes clusters from characters in a document by perceptual grouping process based on Gestalt laws. The concept-driven process applies hypotheses to text clusters to solve incorrect perceptual grouping. The interaction between concept-driven process and data-driven processes verifies hypotheses and forms a layout structure of a document consisting of the hierarchical relationships between text blocks and text lines. Experimental results obtained for 150 documents show the method is adaptable to various layout styles and is effective for complex documents.

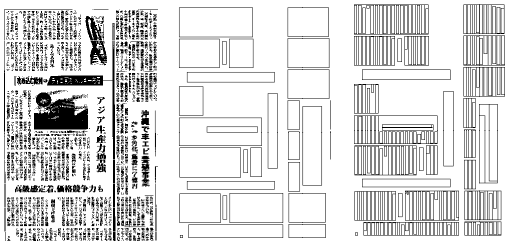
1. はじめに

インターネットとWWWの爆発的普及と，通信技術およびコンピュータ処理能力の急速な進歩と，記憶媒体の大容量化により，世界中で公開されている情報に容易にアクセスできるようになった．これにともない過去の膨大な出版物や現在大量に出回っている印刷媒体をディジタル化する動きが活発になっており，それらを低コストで正確にディジタル化するドキュメントリーダ（文書OCR）に対する要求が高まっている．

印刷文書では紙面の有効活用と記事の一瞥性を実現するために，1次元列のテキスト情報が紙面上で2次元的に割り付けられている．このため文書をディジタル化する際には，レイアウト解析処理（layout analysis）により文書画像から意味のあるテキストのまとまりをレイアウト要素（layout objects，**図1**参照）として抽出したあと，それらを順序付けることにより，最終的に整合のとれた1次元列のテキスト情報に変換する必要がある．このようにレイアウト解析は文書読み取りタスクにおいて重要な役割を果たしており，これまでに多くの研究がなされている．

従来のレイアウト解析研究は概念駆動型方式（concept-driven processing）とデータ駆動型方式

† 株式会社東芝研究開発センター
Research and Development Center, Toshiba Corporation



(a) Document image (b) Layout objects (Text blocks) (c) Layout objects (Text lines)

図 1 レイアウト解析の例
Fig. 1 Example of document layout analysis.

（data-driven processing）に大別することができる．概念駆動型方式は，特定の文書を対象としたモデル/知識を構築し，それを文書画像に適用することでレイアウト要素をトップダウンに抽出するものである．モデルの記述方法としてルールを用いるもの[1),2)]，フレームを用いるもの[3),4)]，AND/OR 木を用いるもの[5)]，意味ネットワークを用いるもの[6)]などがある．概念駆動型方式はノイズの影響に頑健であり，複雑なパターンを対象とする際にも必要最低限の処理で済むという利点がある．しかし多様な印刷文書を処理対象とする場合には，適用すべきモデルの予測を誤ると大きな処理誤りを犯すという危険性がある．また，モデルや知識の構築に大きな負担が生じるという問題がある．

データ駆動型方式は文書画像から特徴をボトムアップに抽出・構造化し，それらを評価することでレイアウト要素を抽出する帰納的で着実な方法である．従来方式としてテクスチャ解析を用いて領域分割を行うもの[7)~9)]，背景分布に基づいてレイアウト要素を分割するもの[10),11)]，文字成分から文字行をボトムアップに構成するもの[12)~14)]などがある．データ駆動型処理では，文書画像がノイズを多く含んでいる場合やレイアウト構造が複雑である場合には特徴抽出誤りが生じて，候補を適切に絞りきれず解釈誤りが生じるという問題がある．

日本語文書では写真/絵/図形などの非文字領域と，複数文字サイズおよび文字フォントで構成される縦書き文章と横書き文章（英文，表，数式を含む）が様々に組み合わされて紙面に配置されている．一般に，縦書き/横書き文章が込み入って複雑に配置されている新聞，大きさの異なる記号を有する数式が紙面の大半を占めている科学技術文献，文字間に大きな空白を持つ文字配置が疎なビジネス文書や名刺（それぞれ**図 2** 参照）などではレイアウト構造の性質がかなり異なるので，多種多様なレイアウト構造を一括して高精度に解析することは難しい課題とされている[15)]．

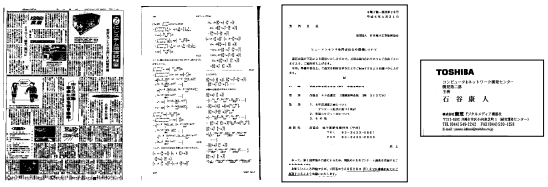


図 2 レイアウト解析が困難な文書の例
Fig. 2 Complex and unconfined layout styles.

本論文では，データ駆動型と概念駆動型の処理が相互作用する知覚循環[16)]により，ノイズ，多義性，情報不足などに起因するデータ駆動型処理の誤りに対して効果的に仮説検証を行う新しいレイアウト解析方式を提案する．データ駆動型処理では文書画像中の文字成分が，ゲシュタルト心理学[17)]でよく知られている類同（similarity），近接（proximity），良い連続（good continuation）の法則に従って群化することで，見出しや本文などの “大域的なレイアウト要素” が創発特徴（emergent feature）として体制化するものとする．概念駆動型処理では，データ駆動型処理により生じた群化誤りを解消するための仮説をレイアウト要素に適用するものとする．

また本方式では「知覚体制化は状況が許す限り，つねにできるだけ良い形にまとまろうとする」というプレグナンツの法則[18)]に基づくことで，文書中のレイアウト要素はできるだけ大域的な領域を占めるように体制化するものとする．すなわち，ビジネス文書や名刺ように文字間距離の大きい文字列や文字間距離が不規則である文字列に対しては，群化において近接あるいは連続の法則の適用を緩和することで体制化を生じやすくする．また，数式のように様々な大きさの記号がレイアウト要素に混在している場合には群化において類同の法則の適用を緩和することで体制化を生じやすくする．

以下本論文では，2 章でレイアウト解析システムの構成を述べた後，3 章でレイアウト構造の記述について述べ，4 章から 6 章でレイアウト解析システムの動作原理について説明する．そして 7 章で多様な文書画像を用いた実験結果を示し，提案方式の有効性を評価する．

2. レイアウト解析システムの構成

レイアウト解析システムを以下の 4 つの基本的な処理で構成する．

領域統合（Region grouping）： 互いに近接し，ほぼ同じ文字サイズを持つ文字成分を統合してクラスタを形成する．あるいは，互いに近接し，文字間距離

と文字サイズがほぼ同じであり，文字行方向が同一のクラスタを統合する．

領域解析（Region analysis）： クラスタをテキストの部分領域と仮定して，文字行方向，文字サイズ，文字間距離などのテキストパラメータを推定し，それらに基づいてクラスタから文字行を抽出する．

領域認識（Region recognition）： クラスタを構成する文字行から文字パターンを切り出したあと文字認識処理を行うことで，各文字パターンの確信度と文字行の確信度を計算する．

領域変更（Region modification）： 文字行配置に関する仮説をクラスタに適用して文字成分の過剰な群化を分解する．また，文字認識結果の確信度に基づいたコンテンツの妥当性に関する仮説をクラスタに適用することで，非文字成分の検出・棄却，テキストパラメータの推定誤りの修正，文字行抽出誤りの解消を行う．

本方式では領域統合/解析/認識をそれぞれデータ駆動型処理と見なし，領域変更を概念駆動型処理と見なしている．レイアウト解析システムでは**図 3** に示すように領域統合処理を最下層とし，その上に領域解析と領域認識を順に配置し，最上位に領域変更を配置する階層構造を採用している．隣接する階層間では双方向の情報通信すなわち相互作用（interaction）を可能としており，最上位の領域変更は他のすべての処理モジュールとの相互作用を可能としている．これにより 4 つの処理は循環して動作するようになっており，処理対象のレイアウト構造に応じてシステム全体の動作を形成することが可能となっている．本論文では 4 つの基本的な処理の組合せにより，ほとんどの印刷文書からレイアウト構造を抽出することが可能であると仮定している．

前処理によって文書画像から抽出された文字成分が初期レイアウト要素としてレイアウト解析システムに入力されると，まず領域統合処理により文字成分の群化が生じて，均質な文字成分で構成されるクラスタが形成される．文字成分の群化が一段落すると，領域解析処理によってクラスタから文字行が抽出され，さらにテキストパラメータが推定されることによりクラスタの幾何的性質が詳細に分析される．そして，領域統合/解析の相互作用によってクラスタの群化が促進され，テキストパラメータが均質であり，文字が規則正しく配置されている大域的な（良い性質を持つ）レイアウト要素が形成されるようになる．このあと各クラスタに対して領域認識処理が適用され，文字認識処理によりレイアウト要素のコンテンツが同定される．こ

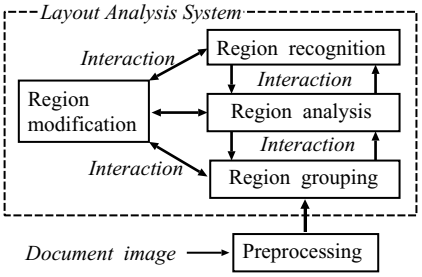


図 3 レイアウト解析システム
Fig. 3 Layout analysis system.

のようにデータ駆動型処理は文書画像から抽出した特徴（文字成分）をボトムアップに処理するようになっており，上位階層のモジュールほどテキストクラスタに対してより詳細な処理を行うようになっている．領域統合→解析→認識→統合→…と処理が繰り返されれば下位階層の処理で上位階層の処理結果が利用されるので，徐々にクラスタの不確定性が減少して，大域的なレイアウト要素が形成されるようになる．

最上位の概念駆動型処理はそれぞれのデータ駆動型処理と通信可能であり，データ駆動型処理における処理誤りを解消するための仮説をクラスタに対してトップダウンに適用するものである．そして，モジュール間相互作用を介してデータ駆動型処理と概念駆動型処理が相補的に協調することにより仮説検証処理が行われ，データ駆動型処理と概念駆動型処理が循環して生起する知覚循環を実現することが可能となる．この知覚循環により，前章で述べたデータ駆動型処理と概念駆動型処理のそれぞれの欠点が効果的に解消され，最終的に正しいテキストクラスタがレイアウト要素として体制化するようになる．

3. レイアウト構造の記述

本章ではレイアウト解析システムに入力される文字成分の抽出（**図 4** (a) 参照）とシステムから出力されるレイアウト構造の記述について説明する．

レイアウト解析処理に先立って，入力文書画像に対して 2 値化処理と傾き補正処理[19]を順次適用する．その結果得られた画像では左上端を原点としている．この画像に対してラベリング処理を施して黒画素の連結成分を抽出する[20]．各連結成分はその外接矩形により表現されるものとする．連結成分は以下の条件式により “文字成分”，“フィールドセパレータ”，“その他” のいずれかに分類される．ただし，RA：連結成分のクラス，RW：連結成分の横幅，RH：連結成分の縦幅，RN：連結成分の総数，$R_{\max} = \max(RW, RH)$，

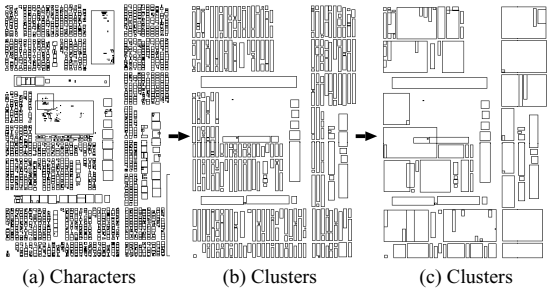

(a) Characters (b) Clusters (c) Clusters

図 4 文字成分の群化の例
Fig. 4 Example of perceptual grouping process.

$R_{\min} = \min(RW, RH)$，$R_{ratio} = R_{\max}/R_{\min}$ とし，κ と ν を定数とし，TH_{ratio}，TH_{len1}，TH_{len2}，TH_{s1}，TH_{s2} を閾値とする．

If $R_{ratio} > TH_{ratio}$ and $R_{\min} < TH_{len1}$ and $R_{\max} > TH_{len2}$ then $RA \equiv$ フィールドセパレータ
else if $R_{\min} > TH_{s1}$ and $R_{\max} < TH_{s2}$ then $RA \equiv$ 文字成分
else $RA \equiv$ その他

$$TH_{s1} = rs_{ave} \times \kappa,\ TH_{s2} = rs_{ave} \times \nu,$$
$$rs_{ave} = \left(\sum_{m=1}^{RN} R_{\min} \right) / RN$$

文字成分の属性を持つ連結成分は初期レイアウト要素と見なされ，レイアウト解析システムに入力される．

レイアウト解析システムが最終的に出力する文書画像レイアウト構造はレイアウト要素の集合 $\mathbf{O} = \{\mathbf{O_i} \mid i = 1, 2, ..., ON_i\}$ で記述される．レイアウト要素 $\mathbf{O_i}$ は以下で定義されるように文字行の集合 $\mathbf{S_i}$，文字成分の集合 $\mathbf{C_i}$，テキストパラメータ $\mathbf{P_i}$ で構成されるものとする．ここで，SN_i と CN_i をそれぞれレイアウト要素中の総文字行数と総文字数とする．

$$\mathbf{O_i} = (\mathbf{S_i}, \mathbf{C_i}, \mathbf{P_i})$$
$$\mathbf{S_i} = \{\mathbf{S_{iu}} \mid u = 1, 2, ..., SN_i\}$$
$$\mathbf{S_{iu}} = (\mathbf{SP1_{iu}}, \mathbf{SP2_{iu}}, SW_{iu}, SH_{iu}, SCF_{iu})$$
$$\mathbf{C_i} = \{\mathbf{C_{ig}} \mid g = 1, 2, ..., CN_i\}$$
$$\mathbf{C_{ig}} = (\mathbf{CP1_{ig}}, \mathbf{CP2_{ig}}, CW_{ig}, CH_{ig}, CCF_{ig}, Code_{ig})$$
$$\mathbf{P_i} = (TD_i, \mathbf{OCS_i}, \mathbf{D_i}, \mathbf{MP_i})$$
$$\mathbf{OCS_i} = (OCW_i, OCH_i),\ \mathbf{D_i} = (HD_i, VD_i)$$
$$\mathbf{MP_i} = (HMP_i, VMP_i)$$

文字行 $\mathbf{S_{iu}}$ は，左上端位置座標 $\mathbf{SP1_{iu}} = (SX1_{iu}, SY1_{iu})$，右下端位置座標 $\mathbf{SP2_{iu}} = (SX2_{iu}, SY2_{iu})$，横幅 SW_{iu}，縦幅 SH_{iu} で記述される文字行パターンの外接矩形データと，文字認識処理時に計算される文字行の確信度 SCF_{iu} で構成される．文字成分 $\mathbf{C_{ig}}$ は，左上端位置座標 $\mathbf{CP1_{ig}} = (CX1_{ig}, CY1_{ig})$，右下端位置座標 $\mathbf{CP2_{ig}} = (CX2_{ig}, CY2_{ig})$，横幅 CW_{ig}，縦幅 CH_{ig} で記述される文字パターンの外接矩形データと，文字認識処理により得られる確信度 CCF_{ig} と文字コード $Code_{ig}$ で表現される．テキストパラメータ $\mathbf{P_i}$ は，文字行方向 TD_i（横書き：0，縦書き：1，不明：-1 とする），文字サイズ $\mathbf{OCS_i}$（$OCW_i = (\sum CW_{ig})/CN_i$，$OCH_i = (\sum CH_{ig})/CN_i$），文字間距離 $\mathbf{D_i}$（HD_i：水平方向文字間距離，VD_i：垂直方向文字間距），統合範囲 $\mathbf{MP_i}$（HMP_i：水平方向統合範囲，VMP_i：垂直方向統合範囲）で構成される．文字行方向 TD_i と文字サイズ $\mathbf{OCS_i}$ はレイアウト要素間の類似性の判定に用いられ，文字間距離 $\mathbf{D_i}$ と統合範囲 $\mathbf{MP_i}$ はレイアウト要素間の近接性の判定に用いられる．

4. 処理モジュールの動作原理

本章では，レイアウト解析システムを構成する 4 つの処理の動作原理について説明する．

4.1 領域統合

レイアウト解析システムに初期レイアウト要素の集合 $\mathbf{O}$ が与えられると，まず領域統合モジュールで近接，類同，連続の法則に基づいたレイアウト要素の群化が生じて，図 4 (b)，(c) のような部分的なクラスタが形成される．群化は，2 つのレイアウト要素 $\mathbf{O_i}$，$\mathbf{O_j}$ を以下の手順により統合して，新しいレイアウト要素 $\mathbf{O_i'}$ を形成することを基本としている．

(1) **レイアウト要素の統合範囲の設定**

レイアウト要素 $\mathbf{O_i}$ に対して水平方向統合範囲 HMP_i と垂直方向統合範囲 VMP_i を次式に従って設定する．

$$HMP_i' = \begin{cases} HD_i + \rho, & HD_i > 0 \\ HMP_i + \rho, & HD_i = 0 \end{cases}$$

$$VMP_i' = \begin{cases} VD_i + \mu, & VD_i > 0 \\ VMP_i + \mu, & VD_i = 0 \end{cases}$$

$$\rho = OCH_i \times \alpha, \quad \mu = OCW_i \times \alpha$$

ここで，HMP_i' と VMP_i' を更新後の統合範囲，α を 1 未満の定数とする．たとえば，水平方向統合範囲 HMP_i' は，レイアウト要素 $\mathbf{O_i}$ の水平方向文字間距離 HD_i にマージン ρ を加えることで**図 5** のような範囲をとり，水平方向に距離 HMP_i' 以内に近接している他のレイアウト要素と統合可能であるとする．

(2) **ゲシュタルト法則に基づく群化**

レイアウト要素 $\mathbf{O_i}$ と $\mathbf{O_j}$ が以下に定義する近接の条件 1 と，類同の条件 1〜2 と，連続の条件 1 をすべて

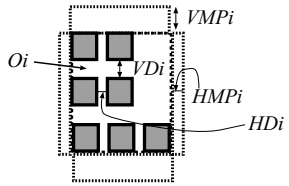


図 5 レイアウト要素の統合範囲
Fig. 5 Object grouping scope.

満たす場合，それらを統合して新しいレイアウト要素 $\mathbf{O_i'}$ を生成する．ここで，

$$
\begin{aligned}
DIST_H(i,j) &= \min_{ij}(\max(CX1_{ig}, CX1_{jq}) - \\
&\quad \min(CX2_{ig}, CX2_{jq})) \\
DIST_V(i,j) &= \min_{ij}(\max(CY1_{ig}, CY1_{jq}) - \\
&\quad \min(CY2_{ig}, CY2_{jq})) \\
OVLP_H(i,j) &= \max_{ij}(\min(CY2_{ig}, CY2_{jq}) - \\
&\quad \max(CY1_{ig}, CY1_{jq}) + 1) \\
OVLP_V(i,j) &= \max_{ij}(\min(CX2_{ig}, CX2_{jq}) - \\
&\quad \max(CX1_{ig}, CX1_{jq}) + 1)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{C_j} = \{C_{jq} \mid q = 1, 2, ..., CN_j\}, \\
&\mathbf{C_{jq}} = (\mathbf{CP1_{jq}}, \mathbf{CP2_{jq}}, CW_{jq}, CH_{jq}, CCF_{jq}, \\
&\qquad Code_{jq})
\end{aligned}
$$

とする．たとえば**図 6** では，$DIST_H(i,j)$ はレイアウト要素を構成する文字成分間の水平方向の距離の最小値 $d1$ となる．

近接の条件 1： $\mathbf{O_i}$ と $\mathbf{O_j}$ が，以下の水平方向の近接条件か垂直方向の近接条件のどちらか一方を満たせば，近接しているものとする．

水平方向の近接条件：$DIST_H(i,j) < \min(HMP_i, HMP_j)$ かつ $OVLP_H(i,j) > 0$

垂直方向の近接条件：$DIST_V(i,j) < \min(VMP_i, VMP_j)$ かつ $OVLP_V(i,j) > 0$

類同の条件 1： $\mathbf{O_i}$ と $\mathbf{O_j}$ は同じ文字行方向を持つものとする．ただし，一方の文字行方向が “不明” であれば，他方の文字行方向は不明，縦書き，横書きのいずれであってもよいとする．

類同の条件 2： 2 つのレイアウト要素 $\mathbf{O_i}$，$\mathbf{O_j}$ の平均文字サイズはほぼ等しい．以下の条件式が満たされれば $\mathbf{O_i}$ と $\mathbf{O_j}$ の平均文字サイズは等しいとする．

$$
\frac{(\mathbf{OCS_i}, \mathbf{OCS_j})}{\| \mathbf{OCS_i} \| \cdot \| \mathbf{OCS_j} \|} \geq 0.8
$$

連続の条件 1： $\mathbf{O_i}$ と $\mathbf{O_j}$ はフィールドセパレータをまたいで統合されないものとする．

(3) **新しいレイアウト要素の生成**

以下の式に基づいて，$\mathbf{O_i}$ と $\mathbf{O_j}$ を統合した後で，$\mathbf{O_i}$ を更新して新しいレイアウト要素 $\mathbf{O_i'}$ を生成し，$\mathbf{O_j}$ を無効にする．

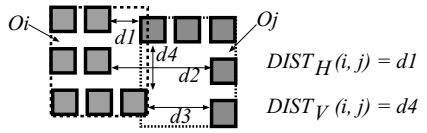


図 6 レイアウト要素間の距離の計算
Fig. 6 Calculation of distance between layout objects.

$$
\begin{aligned}
CN_i' &= CN_i + CN_j,\ \mathbf{C_i'} = \mathbf{C_i} \cup \mathbf{C_j} \\
TD_i' &= TD_i \\
OCW_i' &= \frac{OCW_i + OCW_j}{2} \\
OCH_i' &= \frac{OCH_i + OCH_j}{2} \\
HD_i' &= \frac{HD_i + HD_j}{2},\ VD_i' = \frac{VD_i + VD_j}{2} \\
HMP_i' &= \frac{HMP_i + HMP_j}{2} \\
VMP_i' &= \frac{VMP_i + VMP_j}{2}
\end{aligned}
$$

(4) **反復処理**

手順 (1)，(2)，(3) を新たなレイアウト要素が生じなくなるまで繰り返す．

4.2 領域解析

領域解析モジュールでは，まずレイアウト要素 $\mathbf{O_i}$ の文字行方向 TD_i を検出したあと，文字行方向に隣接する文字成分を統合することにより文字行 $\mathbf{S_i}$ を抽出する．そのあと $\mathbf{S_i}$ から文字サイズ $\mathbf{OCS_i}$ と文字間距離 $\mathbf{D_i}$ をそれぞれ推定する．以下に各パラメータの推定方法と文字行抽出方法について説明する．

(1) **文字行方向の推定**

$\mathbf{O_i}$ を構成する文字成分集合 $\mathbf{C_i}$ に対して，水平方向の文字並びの度合い η_h と垂直方向の文字並びの度合い η_v を計算し，文字並びの度合いが大きい方向を TD_i として採用する．本方式では，文字配置の整列性と連続性に注目した文字並びの度合いを採用することにより，文字間距離 $\geq$ 行間距離である場合を含む多様なレイアウト要素の文字行方向の検出を可能としている[21]．

$$
TD_i = \begin{cases} 0, & \eta_h \geq \eta_v \text{ かつ } diff_{hv} > TH_{CA} \\ 1, & \eta_h < \eta_v \text{ かつ } diff_{vh} > TH_{CA} \\ -1, & \text{その他} \end{cases}
$$

$$
\eta_h = \sum_{N_{\xi h}} \frac{\xi_h}{\min(CH_{ig}, CH_{i(g+1)})}
$$

$$
\eta_v = \sum_{N_{\xi v}} \frac{\xi_v}{\min(CW_{ig}, CW_{i(g+1)})}
$$

$$\xi_h = \min(CY2_{ig}, CY2_{i(g+1)}) - \max(CY1_{ig}, CY1_{i(g+1)}) + 1$$
$$\xi_v = \min(CX2_{ig}, CX2_{i(g+1)}) - \max(CX1_{ig}, CX1_{i(g+1)}) + 1$$
$$diff_{hv} = \eta_h - \eta_v,\ diff_{vh} = \eta_v - \eta_h$$

$N_{\xi h}$，$N_{\xi v}$ を隣接文字成分ペアの総数，TH_{CA} をしきい値とする．ただし，文字成分 $\mathbf{C_{ig}}$ と $\mathbf{C_{i(g+1)}}$ は，$TH_{ovlp} = 0$ とした場合の連続の条件 2 と連続の条件 3 をともに満たす隣接文字成分ペアであるとする．

連続の条件 2：

η_h の場合：$\xi_h > TH_{ovlp}$ かつ $CX1_{ig} < CX1_{i(g+1)}$

η_v の場合：$\xi_v > TH_{ovlp}$ かつ $CY1_{ig} < CY1_{i(g+1)}$

連続の条件 3： 左上端座標 $(TX1, TY1)$ と右下端座標 $(TX2, TY2)$ で構成される矩形領域に他の文字成分が存在しない．ただし η_h の場合，$TX1 = CX2_{ig} + 1$，$TY1 = \min(CY1_{ig}, CY1_{i(g+1)})$，$TX2 = CX1_{i(g+1)} - 1$，$TY2 = \max(CY2_{ig}, CY2_{i(g+1)})$ とし，η_v の場合，$TX1 = \min(CX1_{ig}, CX1_{i(g+1)})$，$TY1 = CY2_{ig} + 1$，$TX2 = \max(CX2_{ig}, CX2_{i(g+1)})$，$TY2 = CY1_{i(g+1)} - 1$ とする．

この連続の法則 2 と 3 では，式中の位置座標値を適切に設定することによって，2 つの文字行（たとえば $\mathbf{S_{iu}}$ と $\mathbf{S_{i(u+1)}}$）の連続性や，2 つの異なる成分（たとえば $\mathbf{C_{ig}}$ と $\mathbf{S_{iu}}$）の連続性を判定することができる．

（2） **文字行の抽出**

文字成分集合 $\mathbf{C_i}$ において，文字行方向に隣接する文字成分を順次統合していくことにより文字行 $\mathbf{S_{iu}}$ を形成する．このとき 2 つの文字成分 $\mathbf{C_{ig}}$ と $\mathbf{C_{i(g+1)}}$ は，連続の条件 1〜3（ただし $\mathbf{O_i}$ が横書きの場合 $TH_{ovlp} = \min(CH_{ig}, CH_{i(g+1)})/3$ とし，縦書きの場合 $TH_{ovlp} = \min(CW_{ig}, CW_{i(g+1)})/3$ とする）と，以下で定義する連続の条件 4 をともに満たしている必要がある．この連続の条件は，同一文字行内では隣接する 2 つの文字 $\mathbf{C_{ig}}$ と $\mathbf{C_{i(g+1)}}$ はある程度近接していることを前提とするものである．

連続の条件 4：

横書き：$\pi_h < (OCH_i \times 1.8)$

縦書き：$\pi_v < (OCV_i \times 1.8)$

$$\pi_h = (\max(CX1_{ig}, CX1_{i(g+1)}) - \min(CX2_{ig}, CX2_{i(g+1)}))$$
$$\pi_v = (\max(CY1_{ig}, CY1_{i(g+1)}) - \min(CY2_{ig}, CY2_{i(g+1)}))$$

文字行 $\mathbf{S_{iu}}$ は，$\mathbf{C_{ig}}$ と $\mathbf{C_{i(g+1)}}$ を用いて，以下の式に基づいて形成される．

$$\mathbf{SP1_{iu}} = (SX1_{iu}, SY1_{iu})$$
$$\mathbf{SP2_{iu}} = (SX2_{iu}, SY2_{iu})$$
$$SX1_{iu} = \min(CX1_{ig}, CX1_{g+1})$$
$$SY1_{iu} = \min(CY1_{ig}, CY1_{g+1})$$
$$SX2_{iu} = \max(CX2_{ig}, CX2_{g+1})$$
$$SY2_{iu} = \max(CY2_{ig}, CY2_{g+1})$$
$$SW_{iu} = SX2_{iu} - SX1_{iu} + 1$$
$$SH_{iu} = SY2_{iu} - SY1_{iu} + 1$$

$\mathbf{S_{iu}}$ が生成されれば，$\mathbf{S_{iu}}$ と隣接する文字成分あるいは新たに生じた文字行との間で同様の統合処理を行う．以上の処理を新しい統合が生じなくなるまで繰り返すことで，$\mathbf{O_i}$ から文字行集合 $\mathbf{S_i}$ を抽出する．このとき，横書き文字行では $SY1_{iu}$ の昇順に $\mathbf{S_{iu}}$ をソートし，縦書き文字行では $SX2_{iu}$ の降順に $\mathbf{S_{iu}}$ をソートすることで，読み順に応じた文字行の順序付けを行う．

（3） **文字サイズの推定**

文字行集合 $\mathbf{S_i}$ から，以下のように文字行幅の平均値 SW_{ave} を計算することで，レイアウト要素 $\mathbf{O_i}$ の文字サイズ $\mathbf{OCS_i} = (OCH_i, OCW_i)$ を推定する．

$$SW_{ave} = \begin{cases} (\sum_{u=1}^{SN_i} SH_{iu})/SN_i, & TD_i = 0 \\ (\sum_{u=1}^{SN_i} SW_{iu})/SN_i, & TD_i = 1 \end{cases}$$

$$OCW_i = \begin{cases} (CW_{ave} + SW_{ave})/2, & TD_i = 0 \\ SW_{ave}, & TD_i = 1 \end{cases}$$

$$OCH_i = \begin{cases} SW_{ave}, & TD_i = 0 \\ (CH_{ave} + SW_{ave})/2, & TD_i = 1 \end{cases}$$

$$CW_{ave} = \frac{\sum_{g=1}^{CN_i} CW_{ig}}{CN_i}$$
$$CH_{ave} = \frac{\sum_{g=1}^{CN_i} CH_{ig}}{CN_i}$$

（4） **文字間距離の推定**

文字行集合 $\mathbf{S_i}$ から文字行間距離の平均値 SD_{ave} を計算し，以下の式を用いてレイアウト要素 $\mathbf{O_i}$ の文字間距離 $\mathbf{D_i} = (HD_i, VD_i)$ を推定する．ここで，$\delta_{u(u+1)}$ を $\mathbf{O_i}$ における隣接文字行の組 $(\mathbf{S_{iu}}, \mathbf{S_{i(u+1)}})$ の総数とし，$\varepsilon_{g(g+1)}$ を $\mathbf{O_i}$ における隣接文字成分の組 $(\mathbf{C_{ig}}, \mathbf{C_{i(g+1)}})$ の総数とする．このとき，$(\mathbf{S_{iu}}, \mathbf{S_{i(u+1)}})$ と $(\mathbf{C_{ig}}, \mathbf{C_{i(g+1)}})$ は連続の条件 1〜3 を満たしているものとする．ここで，CD_{ave} を平均文字間距離，$SD_{u(u+1)}$ を $(\mathbf{S_{iu}}, \mathbf{S_{i(u+1)}})$ の文字行間距離，$CD_{g(g+1)}$ を $(\mathbf{C_{ig}}, \mathbf{C_{i(g+1)}})$ の文字間距離とする．

$$HD_i = \begin{cases} CD_{ave}, & TD_i = 0 \\ SD_{ave}, & TD_i = 1 \end{cases}$$

$$VD_i = \begin{cases} SD_{ave}, & TD_i = 0 \\ CD_{ave}, & TD_i = 1 \end{cases}$$

$$SD_{ave} = \frac{\sum_{v=1}^{\delta_{u(u+1)}} SD_{u(u+1)}}{\delta_{u(u+1)}}$$

$$CD_{ave} = \frac{\sum_{w=1}^{\varepsilon_{g(g+1)}} CD_{g(g+1)}}{\varepsilon_{g(g+1)}}$$

$$SD_{u(u+1)} = \begin{cases} SY1_{i(u+1)} - SY2_{iu}, & TD_i = 0 \\ SX1_{iu} - SX2_{i(u+1)}, & TD_i = 1 \end{cases}$$

$$CD_{g(g+1)} = \begin{cases} CX1_{i(g+1)} - CX2_{ig}, & TD_i = 0 \\ CY1_{i(g+1)} - CY2_{ig}, & TD_i = 1 \end{cases}$$

4.3 領域認識

領域認識モジュールではまず，文字行 $\mathbf{S_{iu}}$ と文字成分 $\mathbf{C_{ig}}$ を用いて文字行ごとに文字パターンを収集することで，レイアウト要素 $\mathbf{O_i}$ から文字行画像を抽出する．次いで，文字行画像に対して文献 22) に基づく文字切り出し/認識処理を適用する．この方式は，文字行中に分離文字（separable characters）や接触文字（touching characters）が混在していても，文字切り出し位置の候補を高精度に推定することができ，文字切り出し位置候補の共起性と排他性を調べることにより複数通りの文字行の仮説を生成することができる．この複数通りの文字行の仮説は複合類似度法[23]に基づいた文字認識処理により検証され，文字認識類似度（similarity）の総和が最大となる切り出しパスを探索することにより，高精度な文字認識結果が得られる．

各文字パターンの文字認識類似度は -1 以上 1 以下の実数値をとり，これを文字の確信度 CCF_{ig} として採用する．認識結果として得られる文字コード情報は $Code_{ig}$ に格納される．また，各文字行で文字の確信度の平均値を計算することで文字行の確信度 SCF_{iu} を計算する．

文字認識処理の結果，文字行を構成するそれぞれの文字の正確な位置，大きさ，順序（読み順），確信度，コード情報を得ることができるので，それらに基づいて文字成分情報 $\mathbf{C_{ig}}$ を書き換え，文字集合 $\mathbf{C_i}$ を再構成する．

4.4 領域変更

データ駆動型処理である領域統合/解析/認識の相互作用により文字成分の群化は一定の段階に達する．しかし，群化により生じたクラスタではテキストパラメータの推定誤りや過剰な体制化が生じている可能性がある．そこでこのようなデータ駆動型処理の誤りを回避するために，領域変更処理ではテキストとしてのコンテンツの妥当性に関する仮説と，文字行配置の適切さに関する仮説をそれぞれのクラスタに対してトップダウンに適用する．具体的には，文字認識確信度に基づいて非文字成分を棄却し（**図 7**），文字行方向検出誤りや文字行抽出誤りを検出・修正する（**図 8**，**図 9**）．また，レイアウト要素における文字行配置の近接性と直列性を仮定して，過剰に統合されたクラスタを分割することで，領域統合による群化誤りを解消する（**図 10**）．以下にそれぞれの処理手続きを説明する．

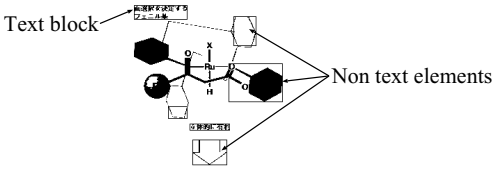


図 7 非文字成分の棄却
Fig. 7 Elimination of non-text elements.

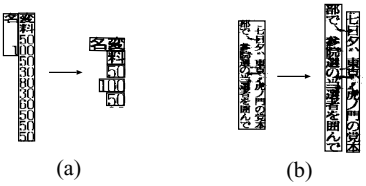


図 8 文字行抽出誤りの修正
Fig. 8 Modification of text line extraction.

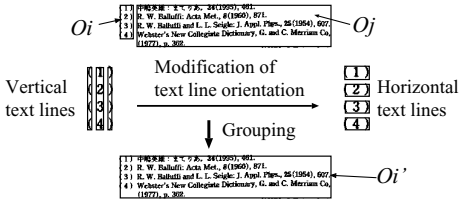


図 9 文字行抽出誤りと群化誤りの修正
Fig. 9 Modification of text line extraction and object grouping.

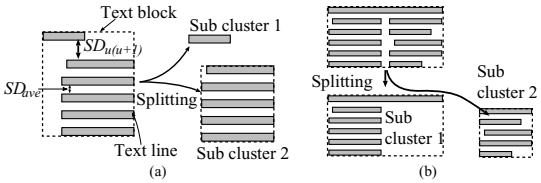


図 10 過剰群化の回避
Fig. 10 Modification of perceptual grouping errors.

（1） 非文字成分の棄却

レイアウト要素 $\mathbf{O_i}$ の文字行 $\mathbf{S_{iu}}$ が $SCF_{iu} < TH_{sim1}$（閾値）を満たす場合，その文字行を非文字成分と見なして棄却したあと $\mathbf{O_i}$ を変更する．$\mathbf{O_i}$ のすべての

文字行が非文字成分と判断された場合には $\mathbf{O_i}$ 全体を非文字成分として棄却する．

（2） **文字行方向の変更**

レイアウト要素 $\mathbf{O_i}$ の文字行 $\mathbf{S_{iu}}$ が $SL_{iu} < TH_{SL}$（閾値）と $SCF_{iu} < TH_{sim2}$（閾値，ただし $TH_{sim1} < TH_{sim2}$ とする）を満足する短い文字行で構成されている場合（図 8 (a)），$\mathbf{O_i}$ の文字行方向が誤りであると仮定して $TD_i' = TD_i \hat{}\ 1$（排他的 OR 演算）により文字行方向を変更する．SL_{iu} は文字サイズで正規化した文字行長であり，以下の式で計算される．

$$SL_{iu} = \begin{cases} SW_{iu}/OCH_i, & TD_i = 0 \\ SH_{iu}/OCW_i, & TD_i = 1 \end{cases}$$

また，図 9 のようにレイアウト要素 $\mathbf{O_i}$ に対して文字行方向が異なるレイアウト要素 $\mathbf{O_j}$ が近接しており，以下のすべての条件式が成り立つ場合，$\mathbf{O_i}$ の文字行方向が誤りであると仮定して文字行方向を変更する．ただし，TH_{dist}：閾値，λ：定数とする

$$OVLP_H(i,j) > 0 \text{ または } OVLP_V(i,j) > 0$$
$$\max(DIST_H(i,j), DIST_V(i,j)) < TH_{dist}$$
$$(OW_i \times OH_i \times \lambda) < (OW_j \times OH_j)$$

（3） **文字行抽出誤りの修正**

領域解析処理で推定された平均文字行幅 SW_{ave} と，文字成分集合 $\mathbf{C_i}$ から推定された文字サイズ $\mathbf{OCS_i}$ が大きく異なる場合，行抽出結果が誤っていると仮定する．具体的には，$\mathbf{O_i}$ が $SW_{ave}/CHS_{ave} > 1.8$ を満たせば，文字行接触（たとえば図 8 (b)）が生じていると仮定し，$\mathbf{O_i}$ における文字行抽出結果が誤っているものとする．ただし，TD_i = 横書きのとき $CHS_{ave} = CH_{ave}$ とし，TD_i = 縦書きのとき $CHS_{ave} = CW_{ave}$ とする．この場合，文字サイズ（横書きの場合 CH_{ig} とし，縦書きの場合 CW_{ig} とする）が閾値 $TH_{chs} = CHS_{ave} \times 1.5$ より大きい文字成分を $\mathbf{C_i}$ から取り除いたあと再び $\mathbf{O_i}$ において文字行抽出を行う．このとき，連続の条件 4 が適用されるため，新たに抽出された文字行が取り除いた文字の箇所で途切れてしまう場合には途切れた文字行を統合する．

（4） **過剰体制化の回避**

図 10 (a) のようにレイアウト要素 $\mathbf{O_j}$ において，隣接する文字行間の距離 $SD_{u(u+1)}$ が平均文字行間距離 SD_{ave} よりも目立って大きい場合や，図 10 (b) のように文字行方向に複数の文字行が並列しており，それが整然と連続している場合には過剰な群化が生じたと仮定する．このような群化誤りは，プレグナンツの法則に基づいて小規模なレイアウト要素の群化が促進されることにより生じる（詳細は後述）．図 10 (a) の場合には $\mathbf{O_i}$ から，$SD_{u(u+1)} > SW_{ave} \times 1.8$ と $SD_{u(u+1)} > SD_{ave} \times 1.5$ をともに満たす隣接文字行の組を検出し，その隣接文字行の間で $\mathbf{O_i}$ を分割して 2 つのサブクラスタを作成する．

図 10 (b) の場合には，連続の条件 1〜3 と以下の連続の条件 5 をすべて満たす隣接文字行 $\mathbf{S_{iu}}$ と $\mathbf{S_{i(u+1)}}$ のグループ化を行う．

連続の条件 5：

横書き ： $\phi_h > \min(SW_{iu}, SW_{i(u+1)}) \times 0.8$

縦書き ： $\phi_v > \min(SH_{iu}, SH_{i(u+1)}) \times 0.8$

$$\phi_h = \min(SX2_{iu}, SX2_{i(u+1)}) - \max(SX1_{iu}, SX1_{i(u+1)}) + 1$$
$$\phi_v = \min(SY2_{ig}, SY2_{i(g+1)}) - \max(SY1_{ig}, SY1_{i(g+1)}) + 1$$

このとき，サブクラスタが 2 つ以上存在し，かつ図 10 (b) のように並列していればそれらを分割して新しいレイアウト要素を作成する．

5. モジュール間の相互作用に基づく知覚循環

本章ではレイアウト解析システムを構成するモジュール間の相互作用とその効果について説明する．

5.1 領域統合と領域解析の相互作用

レイアウト解析過程の初期段階では領域統合処理により，類同，近接，連続の条件に基づいた文字成分の群化が生じて，部分的なクラスタが形成される．それぞれのクラスタでは文字サイズが均質で文字行方向が同一であり，文字が連続して配置されていると見なすことができる．したがって領域解析処理では文字行方向を同定することにより，その方向に隣接する文字を統合するという文字配置の一様性を仮定した簡単な処理により，クラスタから文字行を抽出することができる．さらに文字行の幾何情報を用いて文字サイズを推定することができ，文字行配置から文字間距離を高精度に推定することができる．このようにテキストパラメータが推定されればクラスタの幾何的性質が明確になるので，その後の領域統合処理において正確なクラスタの群化が生じるようになり，大域的な領域が形成されるようになる．大域的なクラスタではさらに安定したテキストパラメータが得られるので，領域統合/解析の相互作用の結果，大域的で良い性質を持つクラスタが体制化するようになる．

5.2 領域解析と領域認識の相互作用

文字サイズが未知である文書画像に対して，文字行中における文字接触，分離文字，低品質文字などの混在を考慮しながら文字切り出し/認識処理を正確に行

うことは多くの処理時間を要する難しい問題である．提案方式では，まず領域解析処理によりレイアウト要素から文字行を抽出することにより，個々の文字パターンを同定するための探索空間を大幅に削減する．さらに平均文字サイズ情報を制約として用いることにより接触文字や分離文字（かすれやとぎれをともなう低品質文字を含む）を効率的に切り出す[22]．この結果，文字切り出し/認識処理によりそれぞれの文字の確信度，位置，大きさなどの情報が明確となるので，その後の領域解析処理で文字認識結果を利用することによりテキストパラメータを正確に推定することが可能となる．このあと領域統合処理では，精密化されたテキストパラメータに基づいた群化が生じるので，さらに大域的で性質の良いレイアウト要素が体制化するようになる．

提案方式ではレイアウト要素において文字サイズの一様性を仮定しているので，組文字，倍角文字，半角文字，分離文字などの大きさの異なる文字が文字接触をともなう形で文字行中に多く混在する場合には文字切り出し誤りが生ずることがある．また，未学習フォントのような文字認識確信度が著しく低い文字で構成される文字列でも文字切り出し誤りが生ずる可能性がある．

5.3 領域変更とデータ駆動型処理の相互作用

領域変更処理では，データ駆動型処理の相互作用により形成されたクラスタに対して，非文字成分の棄却，文字行方向検出誤りの修正，文字行抽出誤りの修正に関する仮説を適用する．この結果，領域解析処理ではこれらの仮説に基づいて再度文字行が抽出され，領域認識処理では新しい文字行に対して再度文字認識処理が実施される．このように領域変更→領域解析→領域認識→領域変更→…を繰り返すことによりテキストパラメータの誤りが修正されるので，テキストとして正しいレイアウト要素が得られるようになる．また領域変更処理では，文字行配置の適切さに関する仮説に基づいてクラスタの過剰な群化を分割する．その結果，領域統合処理により適切な体制化が生じるようになり，正確なレイアウト要素が形成されるようになる．このように提案方式では，データ駆動型と概念駆動型の処理が相補的に働き，それらが繰り返されることで，多義性があるレイアウト構造や複雑なレイアウト構造を対象とした場合でも安定した1つの解釈に自動的にすばやくいたる処理過程を形成することができる．

6. プレグナンツの法則に基づいたレイアウト要素の体制化

本章では，大きさの異なる文字や記号が混在するレイアウト要素や，文字間距離が大きいため文字が孤立して配置されているレイアウト要素のように，類同の法則や近接の法則が適用されないために適切な群化が生じない場合の体制化について述べる．まず，以下の条件式を満たすクラスタ（1文字のみで構成されるものを含む）を“小規模レイアウト要素”（small layout object）と定義する．小規模レイアウト要素の同定は領域認識モジュールで実施される．

小規模レイアウト要素の条件式： $\max(SL_{iu}) < \gamma$ かつ $SN_i < \sigma$，γ と σ は定数

さらに領域変更モジュールにおいて，小規模レイアウト要素の統合範囲が次式に従って更新されるものとする．ただし，ψ を1以上の定数とする．これにより，近接の条件1が緩和されることになる．

$$HMP_i' = HMP_i + OCH_i \times \psi$$
$$VMP_i' = VMP_i + OCW_i \times \psi$$

また，小規模レイアウト要素における文字行抽出では連続の条件4が適用されないものとする．以下に，小規模レイアウト要素の体制化に関するいくつかの事例を示す．

6.1 大域的レイアウト要素の触媒作用による小規模レイアウト要素の体制化

図11 に示す例は，章節タイトルが大きな文字間距離 $DIST_H(i,j)$ を持つために近接の法則が適用されず正しい群化が生じない例である．この場合，小規模レイアウト要素 $\mathbf{O_i}$ と $\mathbf{O_j}$ の下部に配置されている大規模なパラグラフ $\mathbf{O_k}$ が触媒となって小規模レイアウト要素の群化が生じるようになる．具体的には，$\mathbf{O_k}$ の上部および下部に図11の“Object grouping scope”のような統合促進範囲を設ける．この範囲内に存在する2つの小規模レイアウト要素 $\mathbf{O_i}$ と $\mathbf{O_j}$ が類同の条件1〜2と，連続の条件1〜3と，以下に定義する近接の条件2（近接の条件1を緩めたもの）をすべて満たす場合，それらを統合して新しいレイアウト要素 $\mathbf{O_i'}$

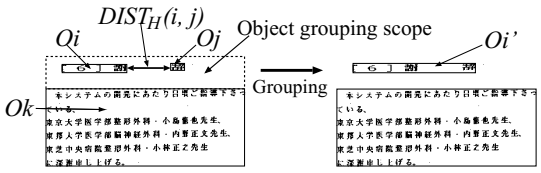


図11 大規模レイアウト要素付近の小規模レイアウト要素の群化
Fig. 11 Perceptual grouping between small objects nearby large object.

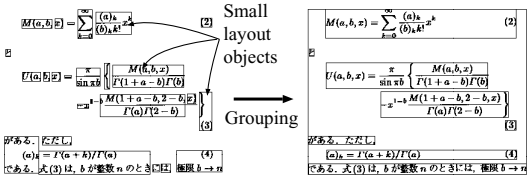


図 12 小規模レイアウト要素の群化
Fig. 12 Perceptual grouping among small layout objects.

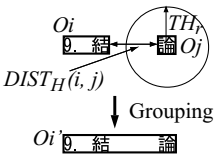


図 13 孤立小規模レイアウト要素の群化
Fig. 13 Perceptual grouping between isolated small objects.

を構成するものとする．

近接の条件 2: $\mathbf{O_i}$ と $\mathbf{O_j}$ は，少なくとも以下の条件式のどちらか一方を満たすこととする．ただし，ω は定数であり，$1 < \omega < 10$ を満たすものとする．

$DIST_H(i,j) < (\max(OCH_i, OCH_j) \times \omega)$ かつ $OVLP_H(i,j) > 0$

$DIST_V(i,j) < (\max(OCW_i, OCW_j) \times \omega)$ かつ $OVLP_V(i,j) > 0$

この結果，近接の条件が大幅に緩められることになるので，図 11 のようにレイアウト要素間距離 $DIST_H(i,j)$ が大きい場合でも体制化が生ずることになる．

6.2 文字サイズが異なる小規模レイアウト要素の体制化

図 12 は，本文の文字サイズよりも大きい数式記号を含む文字が 2 次元的に配置されている例である．これらは本来同じレイアウト要素を構成するはずであるが，文字サイズが異なっており，文字行方向が正しく推定されなかったので類同の条件が満たされず，適切な群化が生じなかった例である．この場合，$\mathbf{O_i}$ と $\mathbf{O_j}$ の統合において類同の法則を適用しないことで群化を促進する．すなわち $\mathbf{O_i}$ と $\mathbf{O_j}$ が近接の条件 1 と連続の条件 1〜3 をすべて満たしていれば体制化が生じるものとする．ただし，$\mathbf{O_i}$ と $\mathbf{O_j}$ のどちらか一方が小規模レイアウト要素であるとする．各小規模レイアウト要素の統合範囲はあらかじめ拡大されているので近接の条件は実質的に緩められた形で適用されることにより，レイアウト要素間の距離がある程度大きくても群化が生じる．さらに，句読点，上付き文字，下付き文字などの微小文字が本文と統合されるようになる．

6.3 孤立小規模レイアウト要素の体制化

図 13 は，章見出しを構成する文字成分が他のレイアウト要素に近接しておらず，文書中に孤立している例である．この場合も文字間距離が大きく，不規則であるため適切な体制化が生じない．そこで，小規模レイアウト要素が以下の孤立性の条件を満たす場合には，それを孤立小規模レイアウト要素と見なす．

孤立性の条件： 小規模レイアウト要素 $\mathbf{O_i}$ の重心から半径 TH_r の円の中に非小規模レイアウト要素が存在しない．

2 つの孤立小規模レイアウト要素 $\mathbf{O_i}$ と $\mathbf{O_j}$ が近接の条件 2 と，類同の条件 1〜2 と，連続の条件 1〜3 をすべて満たせば，群化が生じて新しいレイアウト要素 $\mathbf{O_i'}$ が形成されるものとする．

7. 実験結果

新聞，雑誌，科学技術文献，書籍，ビジネス文書，表形式文書，官報，広告，明細書，レター，名刺など様々なレイアウト構造を持つ印刷文書 150 枚（英文文書を含む）を対象として，本論文で提案したレイアウト解析方式の有効性を評価する実験を行った．文書画像は解像度 16 本/mm のスキャナから 2 値画像として取り込まれたものを用いた．実験システムは PC（CPU のクロック数は 366 MHz）上で C 言語により記述されている．論文中に記載されている種々の閾値や定数などのパラメータは，まずシステム開発時に設計者により直感的に設定され，そのあと約 500 枚の評価対象以外の開発用サンプル画像を用いて試行錯誤的に手動で調整されたことにより得られたものである．実験では，多数の文字で構成されている大域的なレイアウト要素や，文字サイズが均一で文字配置が一様なレイアウト要素に対して文字認識処理を適用しないようにして，実用的な処理速度を実現した．名刺サイズから最大 A3 サイズまでの文書 150 枚の平均処理速度は 2.34 秒/枚であった．実験結果の一例を**図 14** に示す．これらは同一のシステムパラメータで処理されており，本方式が多様なレイアウト構造に対して頑健であることを裏付けている．

実験では，非文字成分の除去，文字行方向の検出，文字行の抽出，テキストブロックの抽出に関する性能を評価した．評価結果を**表 1** に示す．以下では，それぞれの評価項目における処理誤りについて考察する．

（1） **非文字成分の除去誤り**

本方式で用いている文字切り出し/認識エンジンは開発サンプル画像に対して 98.42%の認識性能を達成した高性能なものであり，さらには前述したように文字認識対象を限定したこともあるので，テキスト領域を

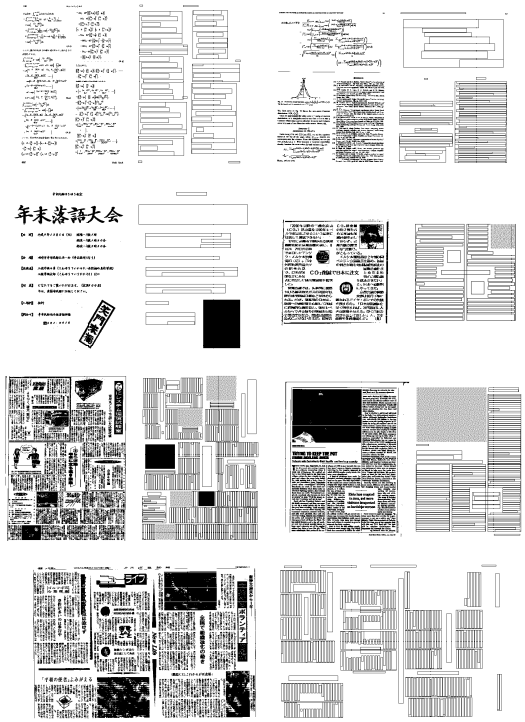


図 14 レイアウト解析結果
Fig. 14 Layout analysis results.

非テキスト領域として棄却することはなかった．しかし，文字認識エンジンはパターンマッチング法に基づいているため，孤立ノイズ，孤立した図の構成要素，小さい図形やロゴ，写真の部分的な塊などが単純な形状をしている場合に高い文字認識確信度を示すことがあり，それらを非文字成分として棄却できないケースがあった．このような問題点を解消するためにパターンマッチング法と構造解析的手法を階層的に実施する方式を採用することが考えられる．

（2） **文字行方向の推定誤り**

文字間距離よりも文字行間距離の方が小さく，かつ文字が格子状に整然と並んでいるテキストブロックにおいて文字行方向の推定誤りが生じた．また，文字行長が極端に短い数字列が多数整列して配置されているケースにおいて文字行方向の推定誤りが生じた．このような処理誤りは，データが空白で区切られている罫線のない表でよく見られた．このような場合には，文字並びの度合いで表現される文字の整列性という評価基準のみで文字行方向を正確に判別することは困難であると思われる．

（3） **文字行の抽出誤り**

文字行方向の検出が誤っていたり，1 行中に組み文字が混在していたり，1 行中に複数行の割注文字行が接触して混在したりしている場合などにおいて文字行抽出誤りが生じた．これは領域解析処理における文字行抽出処理では文字行内の文字配置の一次元性を仮定した単純な方式を採用しており，二次元的な文字行配置に対応していないことが主な原因である．このような場合には，レイアウト要素が体制化したあとで文献 24) のような文字配置の多様性を仮定した詳細な文字行抽出処理を改めて行うことが有効である．

表 1 実験結果
Table 1 Experimental results.

評価項目	成功率（%）
非文字成分の除去	96.7
文字行方向の推定	99.3
文字行の抽出	98.7
テキストブロックの抽出	95.8

（4） **テキストブロックの抽出誤り**

部分的に大きい単語間距離（空白）を持つレイアウト要素や，文字間距離が極端に大きく文字ピッチが均質でないレイアウト要素などにおいてテキストブロックの抽出誤りが見られた．このようなケースでは，本来同じテキストブロックを構成するはずである 2 つの部分的なクラスタの間に大きな空白があり，それが多様なケースとして出現するために本方式で採用した群化規則では対応できなかった．

8. おわりに

多様かつ複雑なレイアウト構造を持つ印刷文書を高精度に読み取ることを目的とした新しい文書画像レイアウト解析方法を提案した．提案方式はデータ駆動型処理である領域統合，領域解析，領域認識と，概念駆動型処理である領域変更で構成されており，処理モジュール間の相互作用によってレイアウト構造の多様性に応じた処理手順の組合せを生成することを可能としている．レイアウト解析過程の初期段階では，領域統合/解析/認識の相互作用により，文字成分が類同，近接，連続の条件に基づいてボトムアップに群化することで，テキストクラスタが形成される．そのあと領域変更処理により，文字行配置やコンテンツの妥当性に関する仮説をクラスタに対してトップダウンに適用することで，過剰な群化やテキストパラメータの推定誤りが解消される．そして概念駆動型処理とデータ駆動型処理が繰り返し行われる知覚循環により，仮説検証が効果的に行われて安定した 1 つの解に高速に到達するようになる．また，文字サイズが不均一なレイアウト要素や文字配置が不規則でスパースなレイアウト要素に対しては，類同性や近接性の基準を緩めること

でレイアウト要素の群化が促進され，最終的に適切な体制化が生ずるようになる．

この結果，横書き/縦書き文章，写真，図表，数式，複数言語，複数文字サイズ，複数文字フォントなどが混在する多様なレイアウト構造を一括して取り扱うことが可能となった．さらに，新聞のように記事が近接して入り組んでいる複雑なレイアウト構造と，名刺やオフィス文書のように文字配置がスパースなレイアウト構造を高精度に解析することが可能となった．

参 考 文 献

1) Esposito, F., Malerba, D., Semeraro, G., Annese, E. and Scafuro, G.: An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization, *Proc. 10th ICPR,* Vol.1, pp.557–562 (1990).
2) Fisher, J.L., Hinds, S.C. and D'Amato, D.P.: A rule-based system for document image segmentation, *Proc. 10th ICPR*, Vol.1, pp.567–572 (1990).
3) Higashino, J., Fujisawa, H., Nakano, Y. and Ejiri, M.: A knowledge-based segmentation method for document understanding, *Proc. 8th ICPR*, Vol.2, pp.745–748 (1986).
4) 黄瀬浩一，杉山淳一，馬場口登，手塚慶一：レイアウトモデルに基づく文書構造解析，電子情報通信学会論文誌，Vol.J72-D-II, No.7, pp.1029–1039 (1989).
5) 駱　琴, 渡邉豊英, 吉田雄二, 稲垣康善, 齋藤隆夫：知識ベースに基づいた図書目録カードの理解，情報処理学会論文誌，Vol.31, No.12, pp.1755–1767 (1990).
6) Bayer, T.A.: Understanding structured text documents by a model based document analysis system, *Proc. ICDAR93*, pp.448–453 (1993).
7) Jain, A.K. and Zhong, Y.: Page segmentation using texture analysis, *Pattern Recognition*, Vol.29, No.5, pp.743–770 (1996).
8) Etemad, K., Doerman, D. and Chellappa, R.: Multiscale segmentation of unstructured document pages using soft decision integration, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.1, pp.92–96 (1997).
9) Tang, Y.Y., Hong Ma, Xiaogang Mao, Dan Liu and Suen, C.Y.: A new approach to document analysis based on modified fractal signature, *Proc.ICDAR95,* Vol.2, pp.567–570 (1995).
10) Antonacopoulos, A.: Page segmentation using the description of the background, *Computer Vision and Image Understanding*, Vol.70, No.3, pp.350–369 (1998).
11) Kise, K., Sato, A. and Iwata, M.: Segmentation of page images using the area Voronoi diagram, *Computer Vision and Image Understanding,* Vol.70, No.3, pp.370–382 (1998).
12) O'Gorman, L.: The document spectrum for page layout analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.15, No.11, pp.1162–1173 (1993).
13) Hönes, F. and Licher, J.: Layout extraction of mixed mode documents, *Machine Vision and Applications*, Vol.7, No.4, pp.237–246 (1994).
14) Gyohten, K., Sumiya, T., Babaguchi, N., Kakusho, K. and Kitahashi, T.: A multi-agent based method for extracting characters and character strings, *IEICE Trans. Information and Systems*, Vol.E79-D, No.5, pp.450–455 (1996).
15) 小川英光（編著）：パターン認識・理解の新たな展開 — 挑戦すべき課題，電子情報通信学会 (1994).
16) Neisser, U.: *Cognition and reality*: *Principles and implications of cognitive psychology*, W.H. Freeman and Company (1976).
17) Spoehr, K.T. and Lehmkuhle, S.W.: *Visual Information Processing*, W.H. Freeman and Company (1982).
18) Koffka, K.（著），鈴木正彌（監訳）：ゲシュタルト心理学の原理，福村出版 (1988).
19) Ishitani, Y.: Document skew detection based on local region complexity, *Proc. ICDAR93*, pp.49–52 (1993).
20) Tsujimoto, S. and Asada, H.: Major components of a complete text reading system, *Proc. IEEE*, Vol.80, No.7, pp.1133–1149 (1992).
21) 石谷康人：文書構造解析のための前処理，電子情報通信学会技術研究報告，PRU92-32，pp.57–64 (1992).
22) Ariyoshi, S.: A character segmentation method for Japanese printed documents coping with touching character problems, *Proc. 11th ICPR*, Vol.2, pp.313–316 (1992).
23) Iijima, T., Genchi, H. and Mori, K.: A theory of character recognition by pattern matching method, *Proc. 1st IJCPR*, pp.50–57 (1973).
24) 宇田明弘，有吉俊二，石谷康人：隣接関係ネットワークに基づく文字列抽出，電子情報通信学会総合大会，D-460, p.248 (1996).

(平成 12 年 11 月 29 日受付)
(平成 13 年 9 月 12 日採録)

石谷 康人（正会員）

1966 年生．1988 年明治大学工学部電気工学科卒業．1990 年明治大学大学院工学研究科電気工学専攻博士前期課程修了．同年,（株）東芝，総合研究所配属．現在，同研究開発センターにて文字認識，文書構造解析・認識，ナレッジマネジメントの研究開発に従事．平成 5 年度電子情報通信学会学術奨励賞受賞．日本ファジィ学会会員．