印刷原稿の多言語識別

Automatic Language Identification in Printed Document Images

大黒　慶久*
Yoshihisa OHGURO

要　　旨

文書画像をスキャンするだけで，その文書が何語で書かれているかを自動識別する技術を開発した．識別対象言語は，欧米系6言語（英仏独伊西蘭）にアジア系2言語（日中）を加えた計8言語であり，文字がほとんど書かれていない原稿（写真や図など）とも区別する．本システムは文書画像を領域識別し，文字画像領域に含まれる各行の言語を識別する．言語識別は，行内矩形（文字構成要素の外接矩形）要素の配置特徴の統計的傾向をもとに，まず欧米系／アジア系に大別する．アジア系言語は，行内矩形要素の配置特徴のさらに詳細な統計傾向によって日本語と中国語とに区別される．欧米系言語は，欧米系言語に共通な文字のみの並びの統計的傾向によって区別される．欧米系／アジア系ともに各種の統計的傾向はN-gram法を利用して表現される．本技術によって，欧州グローバル企業のように多言語が混在した環境においても文書処理を自動化することが可能になる．

ABSTRACT

A system that identifies the language of text automatically in printed document images is developed. The system distinguishes 8 languages: English, French, German, Italian, Spanish, Dutch, Japanese, and Chinese. Photos or figures, the documents which have few text line can be distinguished from character based documents. The system begins by segmenting the page into text lines, then identifies the language for each line. Language identification process distinguishes between Latin and Asian lines at first, based on the statistical tendency of the bounding box sequence in the line. To distinguish between Chinese and Japanese, further detailed tendency is used. To distinguish among 6 Latin languages, the statistical tendency of common character sequence in the line is used. For both Latin and Asian languages, the statistical tendencies are represented by using N-gram method. The language identification technique realizes automatic multilingual document processing system.

* ソフトウェア研究開発本部　画像アプライアンス研究所
Image Appliance Lab, Software R&D Group

1. 背景

近年，経済を中心としたグローバル化に伴い，企業活動において流通する文書も従来の母国語中心から様々な言語に広がりつつある．Web文書やE-mailなど，既に電子化されている文書とともに，紙への印刷による情報の流通・流布は，依然として重要な役割をもつと予想される．

技術的にもコンピュータの多言語対応は進んでおり，商用OSでは多言語の表示や入力が可能なものが一般的になっている．また電子化文書の記述構造として有望視されているXML（eXtensible Markup Language）においても，「言語の指定」機能を担う属性（xml：lang）が規定されており，多言語の電子化文書を扱う環境はハードおよびソフトの両面から整いつつある．

既に文字コードになって存在している文書の使用言語および文字コードを識別することは試みられており，多言語対応ブラウザInternetExplorer，NetscapeNavigatorなどでは，文字コードの自動識別が一部実現されている．文字コードだけでなく電子化文書の使用言語の自動識別に関しても，文字コードの出現頻度を統計的に分析する方法[1][2]，識別対象言語の文字に関するN-gram統計を利用する方法[3][4][5]，識別対象言語において特有な文字（アクセント付き文字を含む）に注目する方法[6]，頻出する特徴的な単語に注目する方法[7]などの研究がある．

一方，紙文書の電子化に際しては画像データを元にしなければならないので，既に電子化されている文書よりも多くの困難が伴う．文字画像データの電子化としてはOCR（光学式文字読取）技術が実用化されているが，一般的なOCRでは，認識対象とする言語を予め指定しなければならない．

文書画像から言語を識別する研究としては，文字行の特徴（水平方向に凹面の数，文字高さの頻度，文字外接矩形の高さのプロファイル，射影など）に基づくもの[8][9][10][11]がある．これらの方法によってアジア系文字行と欧米系文字行とを区別することを目的としているが，複数の特徴に基づく識別条件の調整が難しく，またアジア系文字同士（例えば日本語と中国語）あるいは欧米系文字行同士（例えば英語と仏語）を区別することはできない．

欧米系言語に関しては，文字形状コード（形状によって大分類したもの）[12]，あるいは簡単な特徴抽出に基づく仮想クラス[13]の並び傾向に基づいて，文字画像から使用言語を特定する方法があるが，これらの方法を構造が複雑で文字種類の多いアジア系言語に対してまで拡張することは難しい．

本稿では，欧米系言語とアジア系言語とを言語識別対象とし，文書画像からその使用言語を識別する方法を提案する．提案方法は，まず文書画像を領域識別し，文字領域を抽出する．文字領域においては，黒ラン成分（連続した黒画素部分）を求め，その外接矩形に基づいて文字行を切り出す．文字行では，行内矩形（黒ランの外接矩形）の並び傾向から確率的に欧米系／アジア系に2分する．欧米系行と判断された行に対してはOCRを実行し，文字に変換し，その文字の並び傾向から言語を識別する．アジア系行と判断された行に対しては，行内矩形の並び傾向を，さらに詳細に分析し言語を特定する．いずれの段階においても，各行の特徴は訓練データから確率的に学習され，判断基準には調整や例外条件などのヒューリスティクスをほとんど必要としないことが特徴である．

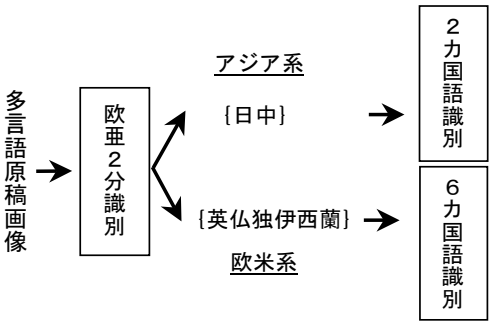


Fig.1　System overview.

以下，本稿では，第2章において提案方式の原理について説明し，第3章では提案方式における実現方法について述べる．第4章は言語識別実験の結果と，それに対する考察であり，第5章では結論とまとめを述べる．

2. 方式の原理

2-1　言語モデル

言語はコミュニケーションの代表的な道具であり，音声言語（聴覚言語）と文字言語（視覚言語）の2種に大別される．諸民族の歴史からも，幼児の言語習得からも，音声言語

が文字言語に先行し，より基本的なものであることがわかる．

音声言語は，時間間隔を持つ聴覚刺激であり，その発生順序に意味がある．つまり音声言語は音響イベントの時系列データとみなすことができる．音声言語における音響イベントに対して概念（文字，シンボルなど，多くは離散的なもの）を対応させ，さらに抽象化することによって成立した文字言語も時系列データとして扱えるであろう．言語とは，コミュニケーションの当時者の間で共有される離散的なシンボルによって成り立ち，構成ルールをもった記号系だといえる．

したがって言語識別という問題は，音声であれ，文字であれ，その時系列な特徴を言語別にモデル化し，識別対象データの時系列特徴と照合するという認識問題に帰着させることができる．

本稿で提案する方法においては，言語モデルとしてN-gramモデルを使用する．N-gramモデルは，情報理論の創始者として知られるクロード・エルウッド・シャノン（Claude Elwood Shannon 1916-2001）が考案した確率的な言語モデルである．言語の生成過程をN-1重マルコフ過程で近似したものであり，「ある言語単位の系列の中で，言語単位のN個の並びの組み合わせが，どの程度出現するか」を調査する言語モデルである．このモデルによって言語の局所的な特徴を表現することができる[14].

以下，N-gramモデルを数学的に定義する．系列W= w_1，w_2，w_3，…，w_nの生起確率P(W)の同時確率は次の条件付き確率の積に分解される．

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1,w_2,\ldots w_{i-2}, w_{i-1}) \quad (2.1)$$

すべての言語単位系列の組み合わせに対して，条件付き確率$P(w_i|w_1, w_2, \ldots w_{i-2}, w_{i-1})$を推定することは現実的には不可能であるから（なぜなら言語現象は固定ではなく，入れ子構造を持ち，無限に生成可能である），これをN-1重マルコフ過程で近似する（一般に，ある事象の確率がその直前のN-1個の事象だけに依存するとき，これをN-1重マルコフ過程と呼ぶ）．Nの値が大きいほど，広い範囲の系列の特徴を表現できるが，系列の組み合わせのバラエティが指数関数的に増大し，実用的には扱いにくくなる．実際にはN=3を使用することが多く，この場合trigramと呼ぶ．trigram(N=3)において，（2.1）式は（2.2）式に書き換えられる．

$$P(W) = \prod_{i=1}^{n} P(w_i|w_{i-2},w_{i-1}) \quad (2.2)$$

$P(w_i|w_{i-2},w_{i-1})$：w_{i-2},w_{i-1}の後にw_iが出現する条件付き確率

$$P(w_i|w_{i-2},w_{i-1}) = C(w_{i-2},w_{i-1},w_i) / C(w_{i-2},w_{i-1}) \quad (2.3)$$

$C(w_{i-2},w_{i-1})$　：系列w_{i-2}，w_{i-1}の出現頻度

$C(w_{i-2},w_{i-1},w_i)$　：系列w_{i-2}，w_{i-1}，w_iの出現頻度

trigramの学習とは，モデル化対象である訓練データに対して$P(w_i|w_{i-2}, w_{i-1})$を求めることに相当し，それには（2.3）式に示すように，3連続の頻度を計数することで算出できる．

定義式から明らかなように，N-gramモデルは，その言語的な単位としては何ら制限するものではなく，文字，単語，品詞など，その言語的意味によらず，離散的なシンボルの時系列として表現可能な対象であれば適用できる．これは言語解析するにあたり，入力が文法的に完結した文である必要はなく，『文の破片』に対しても適用可能なことも意味する．

2-2　文字言語の特徴的な性質

前節では，言語をモデル化する一般的な手段としてN-gramモデルを示した．言語識別するために特徴的な連接傾向をモデル化する言語的単位として何を選択すべきであるかは，識別する対象言語の集合によって異なる．本稿で提案する言語識別は，欧米系言語とアジア系言語とを識別対象とする．

欧米系言語の多くは，同じインド・ヨーロッパ語族に属し，言語間に共通の文字（ローマ字＝ラテン文字）が多数存在する．

一方，いわゆるアジア系言語は，欧米系言語と語族が異り，言語の構成ルールとして，欧米系言語との類似性がなく，使用する文字セットも言語によって異なる．例えば，日本語と中国語は，同じ語族には属さないが，共通に使用する文字（漢字）もある．また日本語は，ひらがな・カタカナという表音文字と，表意文字である漢字とを混在させて表記する．その他，独特の文字体系を使用する言語も多い（ハングル，ミャンマー文字，タイ文字，デーヴァナガリー文字など）．

つまり，欧米系言語は同じ文字セットを使用しているので，文字単独の形状的特徴によって欧米言語間を分類することは困難であり，文字の並び傾向に関する特徴を重視した方式にすべきであろう．したがって，文字そのものを正確に特

定する必要がある．

それに対して欧米系言語とアジア系言語とは使用する文字セットが違うので，文字の形状自体が大きく異なる．よって，文字そのものを正確に特定する必要はなく，両者を区別するに足る，形状的な特徴（＝文字とは限らない）に基づいて，分類することで十分である．

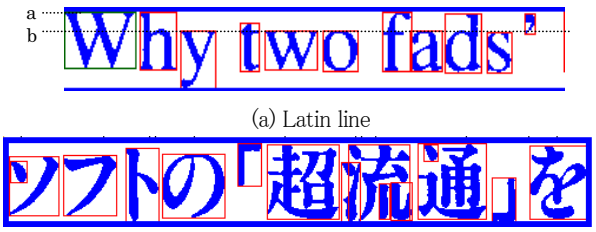
これら文字言語の特徴にしたがって，本稿で提案する言語識別方式では，最初に文字の形状が大きく異なる欧米系とアジア系とを2分し，その後，各言語グループ内で詳細に分類する段階的な分類手段を採用する．

3. 識別アルゴリズム

3-1 欧米系言語とアジア系言語との識別

3-1-1 識別特徴

本稿で提案する言語識別方法においては，アジア系2言語（日本語，中国語）と欧米系6言語（英仏独伊西蘭）の計8言語を識別対象とする．まず最初にアジア系と欧米系とに2分する場合を考える．原稿画像における黒ランの外接矩形を求め，近隣同士を連結していき，行に成長させる．これが行切り出し処理であり，行内に含まれる矩形を行内矩形と呼ぶ．行内矩形は文字の構成要素の外接矩形である．Fig.2に欧米系文字行とアジア系文字行との，行内矩形の配置例を示す．



(a) Latin line

(b) Asian line

Fig.2 Text fragments in English and Japanese document images.

欧米系文字行はFig.2(a)のように，大文字と小文字とが混在していることに加え，アポストロフィー，アクサンテギュ，ウムラウトなど，記号類が存在するので，行内矩形の始点の高さは，Fig.2(a)のaの位置とbの位置との2カ所に集中することは明らかである．一方，アジア系文字行はFig.2(b)のように，漢字，ひらがな，カタカナなど文字の構造が複雑であり，行内矩形の始点の高さは，欧米系文字行で見られるような2カ所への明確な集中はない．

また，アジア系文字行では矩形サイズがバラエティに富むが，欧米系文字行の矩形サイズは数種に限られる．加えて，アジア系文字行（日本語，中国語）では単語間に空白を挿入しないが，欧米系文字行では単語間に空白が挿入される．

さらに，いずれの特徴も隣接矩形との間に関連が見られる．例えば欧米文字行では小矩形が連続することは稀れであることに対して，アジア系文字行においては正方形の矩形が数多く連続することは少ないであろう．

以上のことから，言語の異る文字行を区別するには，注目行において，行内矩形の始点，矩形サイズ（幅，高さ），空白の有無の並ぶ傾向を，言語に特徴的な何らかの基準と比較して判定すればよい．行内矩形は行切り出し処理の過程で既に求まっているので，追加の特徴抽出処理を行う必要がないことも都合がよい．

3-1-2 矩形trigramの導入

ひとつの矩形が行内において，どのように存在しているかを特定する配置情報はFig.3に示すように（始点，高さ，幅）の3成分で表現でき，3次元ベクトルとみなすことができる．つまり行内矩形の時系列情報（左から右への並ぶ順序）は，この3次元ベクトルの系列として扱える．各次元の値を固定段階に量子化することによって，ベクトルのバラエティを有限個に制限することができる．有限個のベクトルの各々にID（ラベル）を付与すれば，行内矩形の時系列情報は，離散的なIDの並び，すなわちシンボル系列に変換することができる．離散的なシンボル系列の特徴は2．1で述べたN-gramを用いて学習することが可能である．

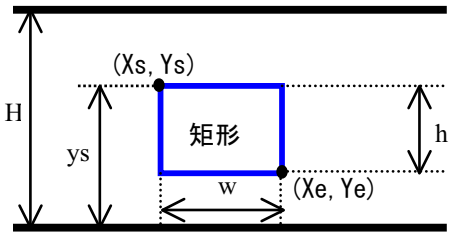


Fig.3 A bounding box in a line.

具体的には以下のように行内矩形の配置情報をシンボル

系列に変換し，これをtrigramによって学習する．行内矩形の（始点，高さ，幅）の3成分を

・矩形の始点の高さ（ys/H）　15段階（4bit）
・矩形高さ（h/H）　8段階（3bit）
・矩形幅（w/H）　2段階（1bit）

に量子化すれば，行内矩形の配置情報ベクトルは240種（= 15×8×2）になり，1byte（=8bits）で表現可能となる．

加えて前述したように，文字行内における空白の有無も当該行の属する言語を特徴づける．行内における空白の存在は，行内矩形の隣接矩形との距離を行高さと比較することによって検出可能である．行高さに対する矩形間距離の割合に，しきい値を設け，しきい値を越えて広く間隔の空いている場合には，空白ありと判定し，空白を意味するベクトルのIDのシンボルを挿入する．

以上の方法によって言語別に訓練用の時系列データを用意すれば，言語に固有な行内矩形の並び傾向を学習できる．Fig.4にtrigram確率表の例を示す．出現確率の演算が高速に行えるよう，確率値の対数を（-1000）倍して整数化した（Score欄）．これによって整数の和算のみで確率値が算出できる．

識別時には，識別したい文字行の行内矩形の配置情報を学習時と同じ要領でシンボル系列に変換した後，学習済みのtrigram表を用いて，その時系列系列の出現確率を言語別に算出し，最も高い出現確率を示す言語を当該文字行の所属する言語であると判断する．一般的には文書における使用言語は原稿毎あるいは領域毎に一種である場合が多く，ある範囲において外来語や固有名詞などを除いて，複数の言語が混在して使われることは稀れである．よって1行毎の識別結果を1枚の原稿全体あるいは領域識別の一領域にわたって集計した後，多数決をとれば当該原稿もしくは当該領域が属する言語を決定することができる．

Trigram	$P(w_i \mid w_{i-2}, w_{i-1})$	Score
$[s_{013}, s_{045}, s_{032}]$	1.0000 (1 / 1)	0
$[s_{013}, s_{064}, s_{033}]$	0.7500 (3 / 4)	287
$[s_{015}, s_{005}, s_{221}]$	0.2307 (6 /26)	1466
$[s_{016}, s_{145}, s_{203}]$	1.0000 (2 / 2)	0
…		

Fig.4　An example of trigram table.

3-2　欧米系言語の詳細識別

3-2-1　欧米系言語における文字セット

本稿で提案する方法において識別対象とする欧米系言語は，西欧諸語（英仏独伊西蘭）とする．これらの言語を表現するには，いわゆる（英語の）アルファベットだけでなく，アクセント付きラテン文字が必要である．以下，各言語で使用される文字種類について簡単にまとめる．

英語　：アルファベット26文字
ドイツ語　：アルファベット26文字に加え，ウムラウト付き文字，エスツェット
フランス語　：アルファベット26文字に加え，綴り字記号がある（アクサン・テギュ，アクサン・グラーヴ，アクサン・シルコンフレックス，トレマ，セディーユ，アポストロフ）
スペイン語　：アルファベット26文字＋1文字（Nの上に～），綴り字記号がある（アセント・アグード・アグード，ディエレシス）
イタリア語　：アルファベット21文字程度（J,K,Q,W,X,Yはほとんど使用されない）
オランダ語　：文法はドイツ語，単語は英語に類似

文字コードとしては，ASCIIコードの空いた部分に補助記号・アクセント付き母音記号を配置した西欧諸語用文字コードISO Latin-1（ISO 8859-1）によってカバーされる．なお，本稿では識別対象とはしないが，中・東欧諸国の言語（チェコ語，スロバキア語，ポーランド語など）を表記するにはISO Latin-2（ISO 8859-2）が用意されている．

3-2-2　欧米系6ヵ国語識別

先述したように，欧米系6ヵ国語は共通する文字（アルファベット）が多いので，文字単独の特徴で区別することは難しい．アルファベットの外接矩形は文字に関わらず類似しており，欧亜2分識別で用いた行内矩形の並びにも言語固有な傾向は明確に表れない．そこで文字そのものを特定し，その並びの傾向に基づいて識別する必要がある．

文字を特定するにあたっては，一つの言語を仮定して欧米言語用OCRによって文字コードに変換する．認識対象となる言語とOCR実行時に仮定した言語とが異なっていれば，

その認識結果には当該言語では使用されない文字が混入することとなり，高い認識精度を得ることはできない．しかしながら，ほとんどの文字は共通であるから，それらの文字が連続する部分については言語に固有な並び傾向が表れる．並びの傾向は欧亜2分識別でも利用したN-gramによって学習する．訓練データは各言語のテキストデータである．この場合，N-gramが学習する時系列データにおけるシンボルは文字そのものに相当する．

3-3　アジア系言語の詳細識別

3-3-1　アジア系言語間の相違点

本稿で提案する方法において識別対象とするアジア系言語は日本語と中国語（簡体）とする．中国語（繁体）は，いわば漢字のオリジナル形であり，日本語中の漢字も中国語（簡体）も，繁体字の一部が各々違う方法で簡略化されたものである．よって中国語（繁体）と，日本語および中国語（簡体）とは，行内の黒画素密度が大きく異なるので，行内の黒画素密度に，しきい値を設けることによって容易に区別可能であり，本提案方式では取り扱わない．

日本語と中国語（簡体）との違いは，日本語は漢字の他に，ひらがな・カタカナを使用する点である．一般的な日本語文における漢字含有率は30%～60%であり[15]，日本語を顕著に特徴づける性質となろう．

3-3-2　アジア系2ヵ国語識別

(1)　ひらがな・カタカナと漢字との相違点

外接矩形の配置特徴に関して，ひらがな・カタカナと漢字との相違点を表現できれば，その尺度を矩形の配置情報ベクトルの次元に追加し，先の矩形trigramと同様に，並び傾向を学習すればよい．

Fig.5に典型的なひらがなとカタカナの矩形抽出結果を示す．図より，次の特徴があることがわかる．

ひらがな・カタカナは，やや小さい矩形が独立し，完全に包含される場合が少ない．

漢字は，大きな正方形の矩形に，小さな矩形が完全に包含される．また，やや大きめの矩形が重複せず，存在する（辺と旁や，冠の場合）．

つまり，矩形が隣接矩形と重複するか否か，包含するか否かの特徴に注目すれば，ひらがな・カタカナと漢字とを区別できる．

先の3次元の特徴は，単独矩形の配置情報のみであり，隣接矩形との位置関係に関する情報は全くない．trigramで矩形の並び情報を学習しても，順序関係は学習できるが，隣接矩形との位置関係（分離，重複，包含）は学習できていない．そこで隣接矩形との距離を特徴として1次元追加し，4次元ベクトルによって矩形の配置情報を表現する．

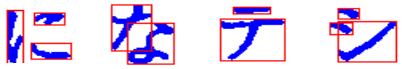


Fig.5　An example of KANA images.

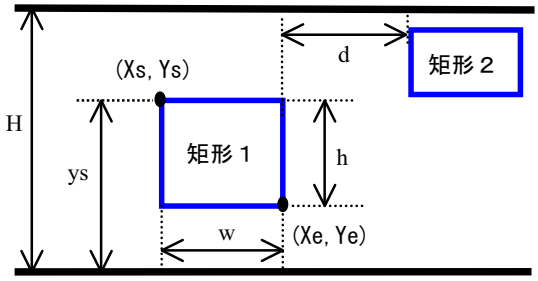


Fig.6　Bounding boxes in a line.

(2)　4次元情報の量子化手法

先の矩形trigramでは3次元ベクトルを量子化する手法として，各次元の値を次元毎に固定段階に量子化し，1byteに格納した．

4次元にすると，1byte内に格納するには量子化段階を少なくする必要があり，矩形分類が大まかになりすぎた結果，制約として機能しにくいことが予想されるので，2bytesに格納する．矩形間距離は隣接矩形が接触した場合に15段階の中間値に収まるよう正値に変換している．

・矩形の始点の高さ（ys/H）　15段階（4bit）
・矩形高さ（h/H）15段階（4bit）
・矩形幅（w/H）　15段階（4bit）
・隣接矩形との距離（d/H）　15段階（4bit）

これによって，生成されるシンボルのバラエティは15の4乗＝50625種となる．ちなみに先の3次元の場合，240種であり，2bytesシンボルにするとシンボル種類が大幅に増加することがわかる．

(3)　ベクトル量子化の導入

シンボル種類のバラエティが増えると，trigram学習が収束するまでに大量のデータが必要になる．学習が収束しないと，入力データを評価する際に正当なパターンを不正なパターンだと判定してしまう恐れがある．さらに(2)の方法では，ベクトル全体としては似ていても各次元の値が異なっていれば違うベクトルとして扱われるので，ベクトルのID番号に対応するシンボルのバラエティ数が増えて学習が収束しにくいだけでなく，テストデータとの照合の際にも類似ベクトルを別種のベクトルと判定してしまう欠点がある．

そこで，ベクトルの各次元の値が完全に一致しなくても，類似するベクトルを選択できるよう，学習時および適用時にベクトル量子化手法を導入する．ベクトル量子化とは，データの存在する特徴ベクトル空間（本稿においては単独矩形の配置特徴のバラエティ）を分割して，カテゴリを代表するテンプレート（代表ベクトル）によって代表させようという近似手法である．

訓練データとなる多数の矩形配置情報から，それらを代表するベクトルを固定数だけ求め登録しておけば（コードブック），その登録IDをシンボルとして利用できる．

この方法によって，訓練データから学習したシンボルの並び傾向とテストデータとを照合する際に，ベクトルの各次元が完全に一致しなくても，ベクトル全体として類似していれば同じシンボルとして扱うことができるようになる．

以上示したように，ある事象がマルコフ過程とみなせるものであり，何らかの手順でシンボルの時系列データへ変換することができるならば，N-gramモデルを適用可能である．

4. 実験と考察

4-1 欧亜2分識別

矩形trigram学習に用いた訓練データとしては各言語4種類の原稿を用いた．欧米系言語に関しては訓練データの矩形数（学習シンボル数）は文字数に近いが，アジア系言語では矩形数は文字数の数倍となる．

実験に用いた訓練および評価原稿は，解像度400[dpi]で読み取った画像から目視によって文字領域のみを抽出し，言語識別処理部には文字画像だけが入力されるようにしている．この文字画像部に対して行切り出し処理を施し，一行毎に分割する．

1つの行内矩形をFig.3のように3成分で定義し，各成分を（15段階，8段階，2段階）に量子化し，1byteに表現した．これに空白と改行を表現するシンボル2種を加え，シンボル種類は242となる．シンボル系列へと変換された訓練データから言語別にtrigram頻度を計数し，言語別にtrigramの条件付き出現確率を算出する．

訓練データのシンボル数とtrigramの種数との関係をFig.7に示す．訓練データのシンボルが少ないうちは，trigram種数は線形に増加するが，訓練データのシンボルが増えるとtrigram種数の増加率は鈍る．

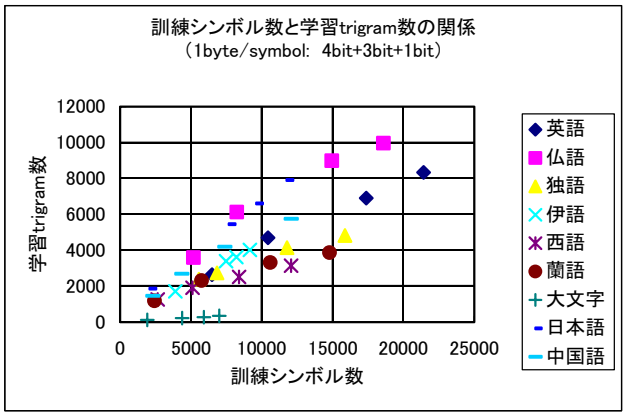


Fig.7 Learning curves of trigram.

評価原稿は各言語40～600枚の原稿を用意した．矩形trigramの訓練時と同じ行切り出し処理を経て，矩形からシンボルへ変換される．本方式は，画像特徴として行内矩形のサイズと位置を用いるので，フォントの違いには影響されない．また文字間ピッチの情報を用いておらず，矩形始点の順序関係のみに着目していることから，固定ピッチ／プロポーショナルピッチにも依存しない．

実験の結果，欧亜2分識別の精度は99.86%（=696/697）であり，評価原稿中1枚のみ，独語原稿を日本語原稿に誤った．この原稿は，大文字のみから構成されるタイトルのような短い行が少数と，小さな文字で書かれた注釈行の少数から構成されており，文字行の言語的な特徴が本質的に現れにくいものであった．

4-2 欧米系6ヵ国語識別

訓練データは各言語4～6種類の原稿から収集した1300～

1600文字のテキストデータである．

評価データは，訓練データとは重複しない，各言語39～180種の原稿に対して欧米系言語用OCR，ExperVision社のRTK[17]を独語に設定にして実行した認識結果テキストである．

Table.1に識別精度をConfusion Matrix形式（混同行列：ある言語がどの言語に識別されたかを縦軸と横軸とで表したもの）で示す．青数字は正解数，赤数字は誤り数である．言語によらず，ほとんど全原稿の言語を識別することができた．行毎の識別傾向は次のようにまとめられる．長い行は，ほとんど間違えることなく言語識別が可能である（一部分で低い評価になっても，他言語が高くなり続けることは稀れだから）．短い行は確率値の差が小さいので誤りやすい（一部分で低い評価になったら，回復しないまま行が終ってしまう）．つまり長い行があるほど安定して識別可能であるといえる．識別実験において，識別を誤った原稿は文字行の少ない広告の類の原稿であった．長い文字行の多い文書に対しては誤識別することはなかった．

Table 1　Confusion matrix for 6 latin language identification.

		識別結果						
	原稿数	英	仏	独	伊	西	蘭	識別率［%］
英	180	180	0	0	0	0	0	100
仏	59	0	59	0	0	0	0	100
独	53	1	0	51	0	0	1	96.22
伊	46	0	0	0	45	1	0	97.83
西	61	0	0	0	0	61	0	100
蘭	39	1	0	0	0	0	38	97.44

4-3　アジア系2ヵ国語識別

訓練データは欧亜2分識別の訓練原稿と同じである．ベクトル量子化手法としてはLBGアルゴリズム[16]を用いた．4-1欧亜2分識別で使用した全訓練データに含まれる矩形122747個を3-3-2(2)の量子化方法にしたがって4次元ベクトルに変換し，コードブックサイズ256にまとめた．

欧亜2分識別と同じく，訓練用データは目視によって文字領域を抽出され，行切り出し処理を施し，1行毎の文字画像へと分割される．この行内矩形をコードブック作成時と同じ量子化手法によって量子化し，さらにこの行内矩形と照合する代表ベクトルをコードブックから探すことによって，シンボル系列へと変換する．結局，各次元のスカラー値を量子化した後，ベクトル値として再度，量子化されることになる．シンボル系列へと変換された訓練データから言語別にtrigram頻度を計数し，trigramの条件付き出現確率を算出する．

評価データは欧亜2分識別と同じ原稿セットである．訓練時と同様の手順を経てシンボル系列へと変換され，行毎に言語別の出現確率を求めることによって，当該行の言語を決定する．行毎の言語決定結果を原稿全体で集計し，多数決によって原稿の言語を決定する．

評価原稿は日本語658枚，中国語62枚である．識別実験の結果，日本語原稿に対しては誤識別なしであった．中国語原稿については1原稿のみ日本語原稿と誤識別した．誤識別した原稿は新聞であり，2値化後の画像には地肌色および紙質が原因のノイズが混入していた．

4-4　全自動8ヵ国語識別

領域識別処理は原稿のレイアウトを分析し，文字領域，図領域，写真領域など，特徴別に分類する処理であり，文書画像の全自動処理には必要な処理である．しかしながら，多様なレイアウトを分析することは高度な知的解釈を必要とし，デザイン性の高い原稿に対して精度良く領域を分類することは困難である．

本稿において提案する言語識別方法は，いずれの局面においてもN-gramモデルに基づいており，入力として完全な文であることを求めていない．よって，領域識別処理が正しい文字領域の取得に失敗したとしても，部分的な文字領域を抽出できれば言語識別が可能である．

今までの実験においては，言語識別処理単独の精度を測定するために，目視によって文字領域を抽出した．これは領域識別処理が文字領域抽出に関しては完全に成功した場合に相当する．実際には領域識別処理において文字領域が正しく抽出される保証はなく，領域の欠落や分割ミスが生じる場合は多い．しかしながら全自動処理を行うにあたっては，領域識別処理は必須である．

以下，自動領域識別処理によって抽出した文字領域を入力とする，欧亜2分識別，欧米系6言語識別，アジア系2言語識別の各処理を統合した，8言語識別処理の識別精度を示す．言語識別は段階的に行われるので，各段階が全て正解である場合に，最終的な識別結果は正解となる．文字行が極端に少

ない原稿は写真･図とみなし，言語識別の対象外とする．

Table.2に識別精度を示す．自動領域識別が文字領域の抽出に失敗したために，文字領域を目視で抽出した場合よりも全体的にやや識別精度は低下している．誤識別した原稿は，いずれもレイアウトが複雑な原稿，あるいは文字行の少ない原稿であった．処理速度は原稿の文字数および言語種に依存するが，いずれの原稿においても3～6[sec/sheet]（AMD Athlon 1.2GHz，メモリ512MB）であった．

Table 2　Confusion matrix for 8 language identification.

		識別結果								
	原稿数	英	仏	独	伊	西	蘭	日	中	識別率［%］
英	180	178	0	0	0	0	0	2	0	98.89
仏	59	0	58	0	0	0	0	1	0	98.31
独	53	0	0	51	0	0	1	1	0	96.23
伊	46	0	0	0	45	1	0	0	0	97.83
西	61	1	0	0	0	60	0	0	0	98.36
蘭	66	4	0	0	0	0	62	0	0	93.94
日	658	0	0	0	0	0	0	658	0	100
中	63	0	0	0	0	0	0	3	60	95.24

5. まとめと今後の展開

本稿では，原稿画像を入力とし，アジア系言語および欧米系言語とをN-gramモデルを用いて言語識別する方法について述べた．N-gramモデルは一次元の離散的なシンボル系列をモデル化する手法として広く使われるが，シンボルへの変換方法を工夫することで，2次元的な特徴をもつ文字行画像データに対しても適用可能な手法であることを示した．実験によって，一般的な文書画像データに対して高速かつ高精度に言語識別できることも確認された．領域識別処理と組み合すことによって全自動に原稿画像の言語識別処理が実現できる．また本方式は訓練データから当該言語の特徴を学習し，人間によるヒューリスティクス・調整をほとんど必要としない．それゆえ識別対象言語の追加が比較的容易である．

今後は言語識別に寄与する新たな矩形特徴を追加し，識別精度の向上を図る．また文書画像検索など，画像特徴としての矩形trigramを他分野へ応用することも試みたい．

参考文献

1) Genitiro Kikui：″Identifying the Coding System and Language of On-line Documents on the Internet″, Proceedings of COLING'96 (1996)

2) 前田亮ほか：″ Web 文書の符合系及び使用言語の自動識別″, 信学論D-II, Vol.J84-DII, No.1 (2001)

3) Schmitt,J.C: ″Trigram-based method of language identification″, U.S. Patent 5,062,143 (1990)

4) William B.Cavnar, John M.Trenkle: ″N-Gram Based Text Categorization″, Proceedings ofthe third Annual Symposium on Document Analysis and Information Retrieval, (1994)

5) T.Dunning：″Statistical Identification of Language ″,Technical report CRLMCCS-94-273. Computing Research Lab, New Mexico State University.(1994)

6) Beesley, Kenneth R :″Language Identifier: a Computer Program for Automatic Natural Language Identification of On-line Text″, Language at Crossroads: Proceedings of 29th Annual Conference of American Translators Association, Oct (1988)

7) IBM：″言語識別処理方法″,特許第2837364号，優先日1994/3/14 (米国)

8) DAR-SHYANG LEE, CRAIG R.NOHL: ″Language Identification in Complex, Unoriented, and Degraded Document Images″, Bell Lab.(1996)

9) C.Y.Suen, et al: ″Categorizing Document Images into Script and Language Classes″, Proceedings of the International Conference on Advances in Pattern Recognition , Plymouth, UK(1998)

10) A. Lawrence Spitz: "Determination of the Script and Language Content of Document Images" PAMI(19), No. 3, pp. 235-245. , March(1997)

11) 尾崎正治，A.Lawrence Spitz: ″多言語文章認識システム：Palace″, 富士ゼロックス　テクニカルレポート,Vol.10 (1995)

12) P.Sibun, J.C.Reynar: ″Language Identification: Examining Issues″, Document Analysis And Information Retrieval (1996)

13) Chew Lim Tan, Sam Yuan Sung, et al: ″Text Retrieval from Document Images based on N-Gram Algorithm″, PRICAI 2000 Workshop August (2000)

14) 中川聖一：″確率モデルによる音声認識″, 電子情報通信学会 (1988)

15) 森田正典：″日本文入力方式と鍵盤方式の最適化″, 信学論, Vol.J70-D, No.11 (1987)

16) T.Linde, A.Buzo, and R.M.Gray, "An algorithm for vector quantizer design,"IEEE Tran. Commun., COM-28, No.1, pp.84-95　(1980)

17) http://www.expervision.com/