

マーカーレス拡張書籍のための2次元文字ブロックインデクシング

宮田 章裕[†] 塩原 寿子[†] 藤村 考[†]

本論文では、文章を読む方向とそれに直交する方向を考慮した2次元のブロックを索引・検索のキーとする2次元文字ブロックインデクシング技術を提案し、書籍内の各位置にデジタルコンテンツへのハイパーリンク設置を可能にするシステム BookEnhancer を紹介する。従来、書籍内にハイパーリンクを設置する際はマーカーを用いる方法があったが、この手法は予め書籍内にマーカーを記載する必要がある。一方、システム上に書籍内のテキストと位置を関連付けておけばマーカーは不要である。すなわち、書籍内を撮影した画像をシステムに送信すれば、システムは画像から OCR (Optical Character Recognition) により抽出したテキストを検索語として位置を特定し、その位置に関連付けられたコンテンツを提示できる。このとき、大量の書籍の中から一意に位置を特定するためには、長く連続するテキストを検索語とする必要がある。ところが、一般ユーザが撮影した画像には OCR 誤認識が約 35% 発生するため、長いテキストには誤認識文字が含まれて正しく検索できないという問題があった。特に部分領域画像の範囲が狭い場合、抽出できる検索語数が少なくこの問題は深刻である。提案手法は少ない文字数で各書籍内位置に固有なパターンを表現できるので、OCR 誤認識が発生する狭領域の書籍内部分画像から一意に位置を特定できる。73,231 文書から 1 文書を特定する検証実験では、提案手法はノイズが無い状態で 99%、ノイズが 33% の状態でも 92% の精度を示し、比較手法を上回ることを実証した。

Two-dimensional Letter Block Indexing Method to Enhance a Book without Markers

AKIHIRO MIYATA,[†] HISAKO SHIOHARA[†] and KO FUJIMURA[†]

We present a text indexing method which takes into account not only the reading direction but also the direction perpendicular to the reading direction. By utilizing this method, we develop a prototype system called “BookEnhancer” which enables users to create hyperlinks in a book without markers. Existing approaches utilized markers to detect hyperlinks in a book, but users were required to print markers on books. On the other hand, markers become unnecessary by a search index of book pages and partial images of books, *i.e.*, the system can detect positions using text extracted from images via OCR (Optical Character Recognition) technology and provide users with position associated digital contents. A long sequence of letters is required to be extracted from images in order to identify the exact position from enormous amounts of book data. However, considering that the average OCR error rate is 35 percent in our experiment, it is highly possible that a long sequence of letters contains misrecognized letters and identification fails. In contrast, our indexing method can extract position specific patterns using fewer letters that can identify the exact position under a situation where the partial image is small and extracted text contains misrecognized letters. The experiment result proves that our method can identify the exact document from 73,231 documents with high accuracies of 99 percent and 92 percent under a situation where error rate is 0 percent and 33 percent, respectively.

1. はじめに

情報の電子化が進んだ現代においても、我々が紙媒体の書籍を情報源として活用する機会が多い。これは、書籍が人間にとって取り扱いやすいメディアであることに起因している。すなわち、書籍には、

- 高速ブラウジングが可能
- パラパラめくることで全体の雰囲気の把握や偶然の情報発見が可能

- 全体の中のどこを読んでいるのかが明確
- 人間にとってなじみ深いメディアであり使用法を教わる必要がない

といった特徴があり¹⁾、今なお利便性の高い情報取得手段として利用されている。書籍の形状・操作感を持つデバイスなどに電子データを表示する手法¹⁾²⁾³⁾が今に至るまで数多く提案されていることから、書籍のユーザビリティの高さが認められていることが分かる。

ところが、書籍は紙媒体であるため、扱えるコンテンツがテキスト、図表、写真などの静的なものに限られる。また、紙面のサイズ、ページ数といった制約もあり、記

[†] 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation



図 1 BookEnhancer 利用時の様子

Fig. 1 A usage example of BookEnhancer

録できる情報量には限りがある。この解決を目指す研究が数多くなされており、デジタルコンテンツへのハイパーリンクを書籍中にマーカーとして記載する方法（2章）が主流であるが、予め書籍内にマーカーを記載しなければならないという問題がある。

一方で、システム上に書籍内のテキストと位置を関連付けて検索可能にしておけばマーカーは不要である。すなわち、書籍内を撮影した画像をシステムに送信すれば、システムは画像から OCR (Optical Character Recognition) により抽出したテキストを検索語として位置を特定し、その位置に関連付けられたコンテンツを提示できる。しかし、この方法では、検索に用いるテキストが短いと位置を一意に特定できず、長いと OCR 誤認識文字がテキストに混入して検索が失敗する危険性がある。特に部分領域画像の範囲が狭い場合、抽出できる検索語数が少なくこの問題は深刻である。

本論文では、OCR 誤認識が発生する狭領域の書籍内部分画像から各書籍位置を一意に特定するために、文章を読む方向とそれに直交する方向を考慮した 2 次元のブロックをキーとする 2 次元文字ブロックインデックスを提案する。また、本技術を用いて書籍内の各位置にハイパーリンク設置を可能にするシステム BookEnhancer を紹介する。インデックス作成済みの書籍であれば、ユーザは書籍内の任意領域をカメラ付携帯電話で撮影するだけでその領域に関連付けられたコンテンツを閲覧できる。これにより、例えば、旅行ガイドブックと風景映像を関連付けるサービスが実現できる (図 1)。

本論文の貢献するところは、2 次元文字ブロックインデックス技術を提案すること、本技術が比較方式よりもノイズに強く精度が高いことを大規模データを用いた実験で実証すること、提案手法を用いたシステム BookEnhancer の実現方式を示すことである。

2. 書籍とデジタルコンテンツ

書籍が持つ制約を超えるために、紙媒体の出版物とデジタルコンテンツを連携させる試みが行われている。Billinghurst らは、書籍のページ中にある黒い枠線で囲

まれたマーカーをヘッドマウントディスプレイ越しに見ると、マーカーが存在する位置に仮想オブジェクトが重畳されて見えるシステムを提案した⁴⁾。Metaio 社のシステムでは、マーカーを含むページを PC に接続したカメラで撮影すると、ページに仮想物が重畳された様子を PC 上で確認できる⁵⁾。

これらの方法により、書籍が扱えるコンテンツの制約、情報量の制約を克服できる。しかし、書籍中のページにマーカーを記載しなければならず、2 つの問題が発生する。1 つ目は、マーカーを記載するタイミングが限定されるという問題である。マーカーが書籍中のページに印刷されている必要があるため、書籍が出版される前にマーカーを記載しなければならない。2 つ目は、マーカーが一定面積を専有してしまうという問題である^{*}。面積あたりで掲載コストが発生する新聞・広告ではこの問題は大きい。また、マーカーを不可視にする試み⁶⁾⁷⁾ が数多いことから分かるように、コンテンツ中にマーカーが存在することは人間にとって不自然である。特に、小説など、文学的・芸術的な書籍の各ページにマーカーが記載されていたら興ざめであり、読者に受け入れられないことは想像に難くない。

これらの制約を克服する方法もいくつか提案されている。ペン状小型カメラとラインマーカーを組み合わせた PaperLink⁸⁾ は、ユーザが書籍中の任意位置にラインを引くと、その領域が撮影されてシステムに蓄積される。ユーザがその領域に関するアクションを定義した後、再度その領域を撮影すると定義済みアクションが実行される。Word Anchor⁹⁾ では、予め書籍中の各単語が登場する位置がデータベースに格納されている。ユーザが専用デバイスで書籍の一部領域を撮影すると、その画像に含まれる単語群を元にデータベースに位置問い合わせが行われ、その位置に関連付けられたコンテンツがデバイス上に表示される。これらの方法を用いると、予めマーカーが記載されていない書籍に対しても、任意位置にコンテンツを関連付けられる。

3. BookEnhancer の概要

3.1 研究目標

我々の研究目標は、書籍が持つ手軽さ・高いユーザビリティはそのままに、そこから格段に多くのコンテンツ・情報を得られるという、新しい情報取得スタイルを“実用レベル”で実現することである。より具体的には、ユーザが書籍の任意位置に注目すると、その位置に関連付けられたデジタルコンテンツを閲覧できる仕組みを想

^{*} QR コードは一辺が 1.5cm 程度ないと、読み取り装置のカメラの焦点が合わずに読み取れないことがあるという。

定している。これを実現するためには、下記 3 つの要件を満たす必要がある。

要件 1) 市販書籍をそのまま用いる：特殊マーカーが記載された専用書籍の使用や、市販書籍に何らかの加工を施す方法は実用的とは言えない。より多くのユーザが手軽に本システムを利用するためには、ユーザが取得した市販書籍がそのまま活用できるべきである。

要件 2) 操作が容易である：情報取得操作が煩雑では、書籍の特徴である“めくる”だけという簡便な操作感を損ねてしまう。ユーザは書籍とデバイスの両方を同時に扱わなければならないので、操作は極力簡単であるべきである。

要件 3) 普及デバイスを用いる：高価な専用デバイスでは、提案する情報取得スタイルを実用化することはできない。安価かつ普及している市販デバイスを用いることが望ましい。加えて、どこにでも持ち運べるという書籍のメリットを損なわぬよう、デバイスは軽量・可搬であるべきである。

3.2 BookEnhancer のコンセプト

上記要件を満たす方法を検討する。なお、以降“位置”とは書籍中における位置を意味し、書籍名、ページ、行、列など、位置を表現する情報を“位置情報”とする。(1) まず、要件 1 のように書籍に手を加えずに、ユーザが注目した位置をシステムが把握する手段として、検索システムの概念が適していると考えた。つまり、各位置の内容と位置情報を関連付けたデータベースを事前に構築し、ユーザが着目した位置の内容をシステムに伝達すると、システムがその内容が存在する位置を特定する仕組みである。(2) 次に、要件 2 の容易な操作方法を考える。ユーザが着目した位置にある内容をシステムに簡単に伝達する手段として、我々はその位置を撮影してシステムに撮影画像を送信する方式が適していると考えた。(3) そして、要件 3 のようにユーザが安価で入手でき、軽量・可搬のデバイスとして、我々はカメラ付携帯電話が適していると考えた。

ここで、(1)(2) を踏まえると、システム実現のためには CBIR (Content-based Image Retrieval) の適用が考えられるだろう。すなわち、事前にシステム上に各位置の画像特徴量と位置情報を関連付けてデータベースに格納し、システムがユーザの撮影した書籍内部分画像の特徴量に適合する位置情報をデータベースに問い合わせる方式である。しかし、クエリとなる書籍内部分画像は、たとえ同じ位置であっても撮影者によって撮影範囲・角度・光条件などが様々であることが想像できる。この条件下では適合画像の検索は困難であるし、検索精度を上げるためには計算コストの増加、データベースの肥大化

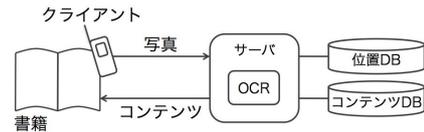


図 2 BookEnhancer のコンセプト
Fig. 2 The concept of BookEnhancer

が避けられない。

我々は発想を変え、この問題をテキスト検索のアプローチで解決する。つまり、事前にシステム上に各位置のテキストと位置情報が関連付けられてデータベースに格納されており、システムはユーザが撮影した書籍内部分画像から OCR で抽出したテキスト断片に適合する位置情報をデータベースに問い合わせる方式である。多くの場合、テキスト検索は CBIR よりも計算コストが低く、必要とするデータ領域も小さい。

上記の議論を踏まえたシステム BookEnhancer のコンセプトを図 2 に示す。クライアントは、写真撮影、通信機能を備えた携帯電話である。位置 DB は、テキストをクエリとし、そのテキストが含まれる位置情報を結果として返すデータベースである。コンテンツ DB は、位置情報をクエリとし、その位置に関連付けられたコンテンツを返すデータベースである。サーバは、ユーザがクライアントで撮影した部分領域画像から OCR でテキスト断片を抽出し、その断片内のテキストを含む位置を位置 DB に問い合わせ、その位置に関連付けられたコンテンツをコンテンツ DB に問い合わせ、得られたコンテンツをユーザ側に送信する。

4. 2次元文字ブロックを用いた書籍内位置索引の提案

4.1 研究課題

前章で述べたコンセプトを実現するためには、書籍内のテキスト断片をクエリとし、その断片が含まれる位置を一意に取得できるような、識別性能が高い検索機能が必要である。もし位置が一意に特定できず複数の候補が検索結果として取得されると、ユーザは候補の中から正しい位置を選択しなければならない、操作の容易性 (3.1 節要件 2) を大きく損ねる。また、できるだけ小さいテキスト断片から検索が行えるべきである。位置特定に必要なテキスト量が少ないほど OCR 処理コストが小さいので処理が高速になるし、書籍内に細かい粒度でコンテンツを関連付けられるからである。検索機能の実現方法としては、文字列が出現する位置を高速検索するために広く使われている転置ファイル方式¹⁰⁾ が妥当と考えられる。例えば、検索対象となる全書籍内のテキストから、N-gram 索引 (索引語が N 文字) による転置ファイルを

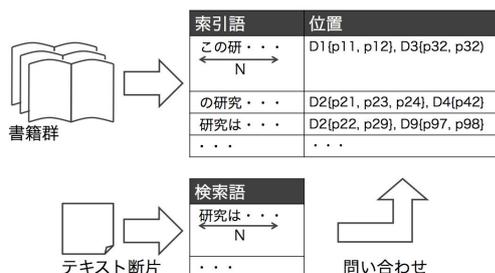


図3 転置ファイル方式を用いた検索
Fig. 3 The concept of inverted index search

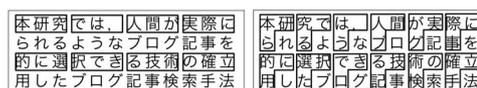
作成する。そして、テキスト断片から N-gram の文字列を検索語として抽出し、この語にマッチする位置情報を取得する (図3)。このとき、検索の識別性能と N は密接な関係にある。文章とは文字がランダムではなく、意味を成すように並んでいる。このため、意味を成す文字の連なりは多くの位置に登場する。つまり、検索の識別性能を高めて位置を一意に特定するためには、索引語・検索語 (以降、キーとする) の文字列単位を長く (N を大きく) し、各位置にできるだけ固有なパターンを抽出する必要がある。例えば、キーが“検索” (N=2) であるより、“検索処理技術” (N=6) である方が検索の識別性能が高く、位置を一意に特定しやすい。

しかし、書籍内の部分領域画像を OCR 処理してテキストデータに変換する際に、OCR 誤認識が発生する点を考慮する必要がある。一般に、商用レベルの OCR 誤認識率は数%程度であるが、これは元画像の撮影状態が良い場合である。BookEnhancer のシステム構成では、片手で書籍を持ち、もう一方の手を持ったカメラ付携帯電話で書籍内の部分領域を撮影することを想定しており、手ぶれ、手の影などが画像の状態を悪化させる。調査実験では、被験者 5 名が撮影した画像を OCR 処理したところ、誤認識率は平均で約 35% であった^{*}。このため、OCR 処理にて得られたテキストデータから文字数が多い検索語を抽出すると、検索語に誤認識文字が含まれて検索が失敗する危険性がある。特に部分領域画像の範囲が狭い場合、抽出できる検索語数が少なくこの問題は深刻である。

そこで、本論文では、**OCR 誤認識が発生する狭領域の書籍内部分画像から各書籍位置を一意に特定できる検索手法の確立**を研究課題として設定する。

4.2 提案手法

前節で述べた研究課題を解決するためには 2 つの条件



(a) キー：連続する N 文字 (b) キー：ブロック

図4 索引語・検索語の抽出方法
Fig. 4 The examples of term extraction

がある。1 つ目は、キーが少ない文字数で構成されていることである。文字数が少なければ、そこに OCR 誤認識文字が混入する可能性は低くなるし、狭領域の部分画像からでも多くのキーを抽出できる。2 つ目は、キーが各書籍内位置に固有なパターンであることである。各位置に固有なパターンであれば、それを利用して位置を一意に特定できる可能性が高い。

そこで我々は、キーを抽出する方法に着目した。通常の転置ファイル方式では、図 4(a) のように文章を読む方向に連続する文字群をキーとしている。一方で、同図 (b) のように文章を読む方向と直交する方向も考慮した 2 次元のブロック単位でキーを作成すると、ブロック内の文字数が少なくても各位置により固有なパターンになると思われる。なぜなら、このブロックに含まれる文字の集合は書籍のページレイアウトや改行位置に起因する“偶然の産物”であるからである。例えば、図 4(a) の枠線内にある読む方向に並んだ 3 文字 (例：本・研・究) よりも、同図 (b) のブロック内にある 3 文字の集合 (例：本・研・ら) の方が、この位置に固有であると考えられる。

そこで我々は、**文章を読む方向とそれに直交する方向を考慮した 2 次元のブロックをキーとする 2 次元文字ブロックインデクシングを提案する。**

5. BookEnhancer の実装

本章では BookEnhancer の実装方法について述べる。図 2 におけるサーバはインデックスモジュールと検索モジュールからなる。クライアントには、NTT ドコモ社の携帯電話 HT-03A を採用した。

5.1 サーバ：インデックスモジュール

インデックス作成処理を図 5 に示す。(1) では、書籍の各ページをスキャンした画像集合とメタデータ (書籍名、ページ番号) 集合を入力する。(2) では、画像を OCR 処理してテキストデータに変換した後、図 4(b) のように L 字型の 2 次元文字ブロック群を抽出する。(3) では、抽出した各ブロックを構成する文字群を連結した文字列 (これを索引語とする) と、そのブロックの位置情報 (書籍名、ページ番号、行、列) を関連付けた転置ファイルを作成し、位置 DB に格納する。表 1 は、図 4(b) のように 3 文字からなる L 字型のブロックを抽出し、位置

^{*} 片手で書籍を持ち、もう一方の手で携帯電話を操作して撮影した画像に基づく結果である。場所はオフィスビル内の照明灯の下、撮影対象は日本語の横書き書籍 (“ユーザが感じる品質基準 QoE” の P.1, P.47, P.119), OCR ソフトウェアは大手メーカーから数万円で販売されているものである。

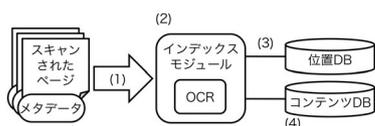


図 5 インデックス作成処理の流れ
Fig. 5 The indexing procedure

情報をブロック内左上端文字の行・列で表現した場合の位置 DB の例である。ブロック内の文字群を別々に扱わずに連結して 1 つの文字列とする理由は、データの問い合わせが文字列をクエリとするシンプルなものになり検索速度を高速化できるからである。上記の処理とは別に (4) では、書籍内の各位置と、そこに設置するコンテンツ (URL など) を表 2 のように関連付けてコンテンツ DB に格納する。

5.2 サーバ：検索モジュール

検索処理を図 6 に示す。(1)では、携帯電話であるページの部分領域を撮影してサーバに送信する。(2)では、画像を OCR 処理してテキストデータを抽出し、そこからインデックス作成処理時と同じ形状 (ここでは 3 文字からなる L 字型) の 2 次元文字ブロック集合を抽出する。図 7(a) は入力画像、図 7(b) は抽出されたテキストデータの例である。一般にユーザが片手で持った携帯電話で撮影した画像では OCR 誤認識が発生しやすい。この例でも、OCR 処理の過程で、2 行目の「る」、3 行目の「択」が形状の似た別の字に誤認識されたとする。(3)では、位置 DB (表 1) に対して、抽出した各ブロックを構成する文字群を連結した文字列 (これを検索語とする) と索引語が一致するレコードの位置情報を問い合わせ、図 8(a) のように位置情報集合を取得する。このとき、誤認識された文字を含む検索語の問い合わせ結果は、偶然一致する索引語が存在する他の位置であったり (同図★印)、一致する索引語が無かったりする (同図☆印)。(4)では、コンテンツ DB (表 2) に対して、各位置情報を領域に含んでいるコンテンツを問い合わせ、図 8(b) のようにコンテンツ集合を取得し、コンテンツごとに件

表 1 位置 DB の例
Table 1 Position DB

索引語	位置情報
本研ら	Book A, Page 2, Line 10, Column 1
研究れ	Book A, Page 2, Line 10, Column 2
究でる	Book A, Page 2, Line 10, Column 3

表 2 コンテンツ DB の例
Table 2 Content DB

コンテンツ	領域
http://1.html	Book A, Page 2, Line 10-15
http://2.html	Book A, Page 3, Line 1-5
http://3.html	Book B, Page 10, Line 14-17

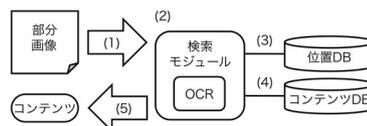
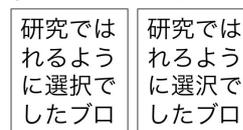


図 6 検索処理の流れ
Fig. 6 The search procedure



(a) 元画像 (b) 抽出結果

図 7 検索処理における元画像・抽出テキストの例
Fig. 7 The source photo and extracted text

研究れ	Book A, P.2, L.10, C.2
★究でる	Book C, P.2, L.3, C.3
ではよ	Book A, P.2, L.10, C.4
れるに	該当なし
...	...

(a) 位置の問い合わせ

Book A, P.2, L.10, C.2	http://1.html	http://1.html 4件
Book C, P.2, L.3, C.3	http://5.html	
Book A, P.2, L.10, C.4	http://1.html	http://5.html 1件
該当なし	...	

(b) コンテンツの決定

図 8 検索結果コンテンツの決定方法
Fig. 8 The procedure of content detection

数を集計して最大件数のもの (ここでは http://1.html) を検索結果と特定する。(5)では、検索結果を携帯電話に送信する。これにより、ユーザは撮影した書籍内の部分領域に設置されていたコンテンツを閲覧できる。

5.3 クライアント

我々は Java でプログラムを開発し HT-03A (Android OS) 上で稼働させた。このプログラムは、撮影画像をサーバに送信し、サーバから返ってくる検索結果コンテンツの URL にアクセスする。OCR 処理や検索処理など計算コストが高い処理はサーバ上で実行するため、現在市販されている程度のスペックの携帯電話であれば軽快に動作する。なお、Android 携帯電話はサードパーティ (我々を含む) が開発したプログラムを簡単にインストールする仕組みが備わっているため、市販の Android 携帯電話にこのプログラムをインストールするだけで BookEnhancer のクライアントになる。これらは実用性を考慮しての判断である。

5.4 システムを利用したサービスイメージ

BookEnhancer の適用対象としては旅行ガイドブックが考えられる。ガイドブック内の各位置に風景映像を関連付けておくと、読者は旅先でガイドブックを広げて次の目的地の様子を事前に把握できる (図 1)。あるいは、教科書内の各位置に講義映像や問題の詳細な解法を関連

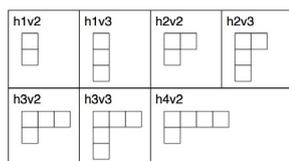


図 9 検証に用いた 2 次元文字ブロックの形状

Fig. 9 2D letter block used in the experiment

付けておき、生徒の自宅学習を支援するというサービスも考えられる。

6. 評価実験

6.1 実験概要

この実験の目的は、BookEnhancer の中核技術である 2 次元文字ブロックインデクシングの有効性の検証である。具体的には、ある小さなテキスト断片を含む位置を検索するタスクを行い、提案手法の (1) 識別性能、(2) OCR 処理誤認識への耐性の検証を行う。提案手法のキー抽出方式は 2~5 文字からなる L 字型 2 次元文字ブロック計 7 方式を検証する (図 9)。文字ブロックを構成する 1 マスは 1 つの文字を表現している。比較手法のキー抽出方式は、N 文字単位で索引語を抽出する N-gram 方式 (N=2~5)、分かち書き*により得られた全単語をキーとする単語方式を検証する。

6.2 実験方法

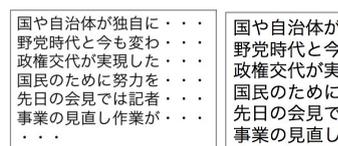
検証用書籍データとして、2008 年 1 月 1 日~12 月 31 日の毎日新聞の全記事 (電子データ版) 計 73,231 件を用いる。1 記事を書籍の 1 ページとみなし、各記事の本文を横書き 1 行 40 文字のテキストファイルに成形し、各記事に ID を付与する。記事の平均文字数は 518 文字、成形後の平均行数は約 13 行である。

次に、インデックスを作成する。提案手法・比較手法の各方式で、全記事の先頭から 1 文字ずつずらしながら索引語を抽出し、抽出元記事 ID と関連付けて位置 DB に格納する。例えば図 10(a) の記事から抽出した索引語は表 3、4 のように格納される。各方式で作成した位置 DB の情報を表 5 に示す。

続いて、テキスト断片を作成する。成形済み記事の中から 7 行以上あるものを無作為に 1000 件抽出し、図 10(b) のように各記事の上から 1~6 行目、左から 1~6 文字目の領域にある計 36 文字を抽出してテキスト断片とする。作成した断片は無作為に 100 件ずつ、10 組の断片データセットに分ける。なお、この断片はテキストファイルから抽出したため、OCR 誤認識文字は含まれていない。そこで、各断片の中から無作為に選んだ最大 50 % の文字をノイズ文字 “_” に置き換える**ことで、

* mecab0.98・mecab-ipadic2.7.0-20070801 を利用した。

** 全記事の中で元々 “_” を含んでいるものは無かった。



(a) 記事 (ID=1000) (b) 断片

図 10 評価実験に用いた記事の例

Fig. 10 The sample of articles used in the experiment

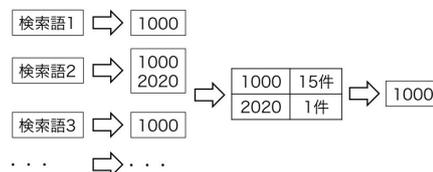


図 11 検索結果決定の処理イメージ

Fig. 11 The determination procedure of search results

誤認識文字をエミュレートする。

上記インデックス、テキスト断片を用いて断片データセットごとに検索タスクを行う。これは、テキスト断片からその断片の抽出元である記事 ID を求めるタスクである。まず、断片データセット内の 1 つの断片から各方式で先頭から 1 文字ずつずらしながら検索語を抽出する。次に、検索語ごとに位置 DB に問い合わせを行い、対応する記事 ID を取得して検索結果候補群に加える。同一検索語に対応する記事 ID が複数ある場合はすべて検索結果候補群に加える。すべての検索語について問い合わせが終了したら、検索結果候補群を集計し、最も多く出現する記事 ID を検索結果とする。図 11 に検索結果決定の処理イメージを示す。この作業をデータセット内の全記事について行う。

6.3 実験結果

表 6 に示すのは、断片データセットごとに正しい検索結果が得られた検索タスクの割合を求め、それを全データセット間で平均した値 (以降、正解率とする) である。ノイズ率とは断片内の全文字数に対するノイズ文字数の割合である。表中の太字部分は、各ノイズ率において最

表 3 評価実験における提案手法の位置 DB の例

Table 3 The position databases of proposal methods

索引語 (h2v2)	位置情報 (記事 ID)	索引語 (h3v3)	位置情報 (記事 ID)
国や野	1000	国や自野政	1000
や自党	1000	や自治党権	1000
自治時	1000	自治体時交	1000

表 4 評価実験における比較手法の位置 DB の例

Table 4 The position databases of baseline methods

索引語 (5-gram)	位置情報 (記事 ID)	索引語 (単語)	位置情報 (記事 ID)
国や自治体	1000	国	1000
や自治体が	1000	や	1000
自治体が独	1000	自治体	1000

良の結果を示した方式の正解率である。

ノイズ率 0 % の場合、提案手法の各方式と N-gram 方式の間には有意な差が見られない。一方で、単語方式の正解率は 0.84 であり、提案手法や N-gram 方式と比べて低い。1 % 水準で t 検定を行うと、提案手法の各方式および N-gram 方式 (N=2~5) はすべて単語方式より正解率が有意に高いことが確認できた^{*}。

ノイズ率が増えると各方式とも正解率は低下する。我々が想定するノイズ率は約 35 % (4.1 節) であることを踏まえ、表 6 のノイズ率 33.3 %, 36.1 % の場合を見ると、h2v2 方式が 0.921, 0.825 と全方式内で最高の正解率を示している。同条件では 3-gram 方式が 0.866, 0.760 と比較手法内で最高の正解率である。各ノイズ率にて h2v2 方式と 3-gram 方式の間で 1 % 水準で t 検定を行うと、どちらの場合も h2v2 方式の正解率の方が有意に高いことが認められた^{**}。

6.4 考察

まず、単語方式について考察する。ノイズ率によらず単語方式は提案手法の各方式、N-gram 方式より正解率が低かった。これには 2 つの理由が考えられる。1 つ目は、位置 DB 内で重複する索引語の多さである。単語方式はユニークな索引語が 0.5 % と少ない (表 5)。すなわち、同一の索引語が多く位置に存在するため、位置を一意に特定できない場合が多かったと推測できる。2 つ目は、テキスト断片の周辺部で単語が途切れる場合に、正しい単語が抽出できないことである。例えば、図 10(b)3 行目では「実現」という単語が途切れいる。また、単語方式はキー抽出時に分かち書きを行う必要があるが、この処理は提案手法や N-グラム方式のように文

字の連続を抽出する処理と比べて計算コストが高く、インデックス作成時、検索時にボトルネックになりうる。これらのことから、単語方式は BookEnhancer の実現には不向きである。

次に、提案手法の各方式と N-gram 方式について考察する。これらは規定文字数のキーを抽出する方式である。利用する文字数が少ないほど、テキスト断片から多くのキーを抽出できるし、検索語にノイズ文字が混入している確率も低い。しかし、キーが短いために位置を一意に特定することは難しい。2 文字を用いる h1v2 方式、2-gram 方式で作成した位置 DB ではユニークな索引語は数%しかなく (表 5)、実験結果においてもノイズ率が増加すると大幅に正解率が低下している (表 6)。逆に、利用する文字数が多いほど、キーが長いために位置を一意に特定しやすい。しかし、文字数が多いために検索語にノイズ文字が混入している可能性が高まり、やはり正解率は低下すると考えられる。ノイズ率を上げると、5 文字を用いる h3v3 方式、h4v2 方式、5-gram 方式は 3~4 文字を用いる方式よりも正解率の低下が大きい (表 6)。

上記の議論を踏まえると、キーの文字数は少なくとも多くても良くない。実験においても 3 文字を利用する h1v3 方式、h2v2 方式、3-gram 方式はノイズ率が増加しても比較的高い正解率を保っている (表 6)。6.3 節で前述のとおり、ノイズ率 35 % 前後では、h2v2 方式が 3-gram 方式よりも有意に高い正解率を示した。また、提案方式内では h2v2 方式が h1v3 方式を上回った。3 方式の中で h2v2 方式が最も高い正解率となった理由は 2 つ考えられる。1 つ目は、h2v2 方式のキーが各位置により固有なパターンであったことである。h2v2 方式のユニークな索引語は 3-gram 方式の 3 倍以上である (表 5)。しかし、ユニークな索引語数であれば、h1v3 方式の方が h2v2 方式よりも多い。そこで考えられる 2 つ目の理由が、テキスト断片から抽出できる検索語の数の違いである。6.3 節で用いた断片は 6 行×6 列であるから、h1v3 方式と 3-gram 方式では検索語が 24 個抽出できる。一方、h2v2 方式はその形状から同じ断片から 25 個の検索語が抽出できるため、より正確に検索が行えた可能性がある。

既存研究としては、Word Anchor⁹⁾ がある。撮影範囲が広く、画像が鮮明な場合は、ページ内における単語の隣接関係を利用して位置を特定する彼らの手法が大変有効である。しかし、本論文で前提としているのは、撮影範囲が狭く、画像が不鮮明な場合である。範囲が狭いと、画像の端で途切れずに抽出できる単語は少ない。画像が不鮮明だと OCR 誤認識により正しい単語が抽出で

表 5 各方式で作成した位置 DB の情報
Table 5 The summary of position databases

	索引語の 総数 (A)	索引語の ユニーク数 (B)	B / A
h1v2	35,035,268	1,880,381	5.4 %
h1v3	32,206,469	19,086,414	59.3 %
h2v2	34,194,093	12,125,120	35.5 %
h2v3	31,435,026	27,660,250	88.0 %
h3v2	33,351,102	21,791,629	65.3 %
h3v3	30,661,822	29,074,866	94.8 %
h4v2	32,506,325	26,551,155	81.7 %
2-gram	36,949,353	484,677	1.3 %
3-gram	35,967,104	3,793,041	10.5 %
4-gram	34,986,767	10,102,853	28.9 %
5-gram	34,008,163	16,457,158	48.4 %
単語	24,474,035	125,733	0.5 %

^{*} Kolmogorov-Smirnov 検定にて各群の正規性、F 検定にて比較する各群間の分散の均一性を確認し、Student の t 検定を行った。
^{**} 同じく各群の正規性を確認したが、各群の分散に均一性が確認できなかったため Welch の t 検定を行った。

表 6 実験結果

Table 6 The evaluation result

ノイズ率	提案手法							比較手法				
	h1v2	h1v3	h2v2	h2v3	h3v2	h3v3	h4v2	2-gram	3-gram	4-gram	5-gram	単語
0.0 %	0.989	0.988	0.990	0.989	0.987	0.989	0.989	0.984	0.980	0.982	0.986	0.838
2.8 %	0.989	0.988	0.990	0.988	0.987	0.988	0.989	0.984	0.980	0.981	0.985	0.800
5.6 %	0.989	0.989	0.990	0.990	0.987	0.989	0.987	0.982	0.978	0.980	0.980	0.745
8.3 %	0.990	0.987	0.990	0.988	0.988	0.989	0.986	0.984	0.979	0.981	0.970	0.643
11.1 %	0.990	0.989	0.988	0.990	0.986	0.985	0.988	0.980	0.980	0.980	0.979	0.497
13.9 %	0.990	0.989	0.989	0.989	0.986	0.987	0.984	0.984	0.980	0.978	0.960	0.277
16.7 %	0.987	0.988	0.988	0.989	0.984	0.981	0.977	0.977	0.977	0.971	0.936	0.115
19.4 %	0.985	0.989	0.987	0.985	0.986	0.980	0.965	0.975	0.976	0.957	0.873	0.038
22.2 %	0.974	0.989	0.990	0.988	0.982	0.962	0.948	0.962	0.973	0.949	0.804	0.011
25.0 %	0.930	0.988	0.985	0.981	0.974	0.943	0.914	0.950	0.962	0.906	0.733	0.004
27.8 %	0.816	0.983	0.987	0.967	0.955	0.891	0.861	0.912	0.962	0.853	0.671	0.002
30.6 %	0.603	0.942	0.967	0.902	0.906	0.822	0.792	0.820	0.932	0.764	0.510	0.002
33.3 %	0.408	0.864	0.921	0.827	0.832	0.721	0.651	0.662	0.866	0.661	0.420	0.002
36.1 %	0.209	0.656	0.825	0.716	0.716	0.600	0.566	0.472	0.760	0.573	0.341	0.003
38.9 %	0.090	0.452	0.655	0.578	0.569	0.478	0.423	0.266	0.620	0.430	0.251	0.001
41.7 %	0.042	0.239	0.425	0.401	0.435	0.345	0.327	0.150	0.435	0.319	0.196	0.001
44.4 %	0.028	0.096	0.227	0.254	0.270	0.259	0.212	0.093	0.224	0.179	0.133	0.001
47.2 %	0.023	0.030	0.102	0.148	0.147	0.161	0.143	0.072	0.130	0.108	0.089	0.001
50.0 %	0.019	0.006	0.046	0.071	0.070	0.092	0.088	0.060	0.064	0.065	0.043	0.002

きない。この状況では単語の隣接関係がうまく検出できない場合がある。例えば、単語 A, B, C の順に並ぶ隣接関係で位置を特定する場合、B に 1 文字でも誤認識文字が含まれると A, B, C のすべてが位置特定に使えない。提案手法では、ブロックの構成単位を単語ではなく文字としているので、画像の端で単語が途切れていてもブロック抽出には影響しない。また、ブロックの位置関係ではなく出現数のみで位置特定を行うため、誤認識文字を含むブロックが領域内に分散していても位置特定が行える。

7. おわりに

本論文では、文章を読む方向とそれに直交する方向を考慮した 2 次元のブロックをキーとする 2 次元文字ブロックインデクシングを提案した。大規模データを用いた検証実験を実施し、OCR 誤認識が発生する狭領域の書籍内部分画像から位置を高精度に特定できることを実証した。さらに、提案手法を用いて書籍から格段に多くの情報を得られるシステム BookEnhancer を実装した。このシステムでは、ユーザは書籍内の部分領域をカメラ付携帯電話で撮影するだけで、その領域に関連付けられたコンテンツを閲覧できる。書籍中にマーカーを記載する必要が無いので、書籍が出版された後からでもコンテンツを登録できるし、コンテンツ登録のためにページ上に余分な領域を占有することもない。また、システム上でインデックス作成済みの書籍であれば、どのユーザの書籍からもコンテンツを閲覧できる。今後の課題は、非等幅フォントへの対応である。現在、非等幅フォントに

対応する機能を開発中である。

参 考 文 献

- 岡田, 松下: 本メディアを超えて: Book Window, 情報処理学会論文誌, Vol.35, No.3, pp.468-477 (1994).
- Card, Hong, Mackinlay and Chi: 3Book: A Scalable 3D Virtual Book, *CHI2004*, pp.1095-1098 (2004).
- 島田, 宇都宮, 鏡原, 田中, 島田, 大倉, 東: 仮想書籍ブラウジングシステムの試作, 情報処理学会論文誌, Vol.46, No.7, pp.1646-1660 (2005).
- Billinghurst, Kato and Poupyrev: The Magic-Book: a transitional AR interface, Vol.25, No.5, pp.745-753 (2001).
- Metaio Inc.: <http://www.metaio.com>.
- Koike, Nishikawa and Fukuchi: Transparent 2-D markers on an LCD tabletop system Export, *CHI2009* (2009).
- Park and Park: Invisible Marker Tracking for AR Export, *ISMAR2004*, pp.272-273 (2004).
- Arai, Aust and Hudson: PaperLink: a technique for hyperlinking from real paper to electronic content Export, *CHI'97*, pp.327-334 (1997).
- 嶺, 亀山, 高橋, 古賀, 緒方: 文字認識と単語レイアウト解析を用いた紙文書とデジタルデータの情報リンク手法, Vol.J92-D, No.6, pp.868-875 (2009).
- Frakes and Baeza-Yates: *Information Retrieval: Data Structures and Algorithms*, Prentice Hall (1992).