社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

1億ページのデータベースを対象とした大規模文書画像検索

竹田　一貴†　　黄瀬　浩一†　　岩村　雅一†

† 大阪府立大学大学院工学研究科
〒 599–8531 大阪府堺市中区学園町 1–1
E-mail: †takeda@m.cs.osakafu-u.ac.jp, †{kise, masa}@cs.osakafu-u.ac.jp

あらまし　本稿では，1 億ページのデータベースを対象とした大規模実時間文書画像検索法を提案する．我々はすでに，1,000 万ページのデータベースから実時間で検索可能な手法を提案している．この手法を用いてさらなる大規模化を実現するためには，より一層のメモリ削減が求められる．同時に，メモリ削減による検索精度の低下を抑制する必要がある．これを実現するため，検索に有効な特徴量のサンプリング法と，特徴量の柔軟な照合を実現するための多重探索法を提案する．1,000 万ページのデータベースを用いた実験から，従来手法と比較して，検索精度を維持したまま 70% のメモリ削減を実現できることが分かった．また，1 億ページのデータベースから，必要メモリ量 236GB，精度 98.7% ，処理時間 26.8ms で検索でき，高いスケーラビリティを持つこと確認した．

キーワード　文書画像検索，文書画像処理，大規模データベース，LLAH

1. はじめに

近年，スマートフォンやタブレット PC の急速な普及に伴い，電子書籍の需要が高まっている．電子書籍は，文字以外に音楽や動画などのデジタルコンテンツを扱えることや，facebook 等の SNS との連携が可能であるという利点を持つ．これにより，文書の持つ表現力を広げ，読者の理解度を向上させることや，読書体験の共有が可能となる．一方，電子書籍への需要が高まりつつある現在でも，紙の文書への需要は低下していない．このような中で，紙の文書からも電子書籍と同様のサービスを享受できることが期待されている．これが実現すれば，世の中に存在するあらゆる文書を “情報を引き出すための窓口” として利用でき，新たなメディアの創出につながると考えられる．

媒体によらないサービスを実現するためには，紙と電子で共通する特徴をキーとして対応する文書を検索する技術が必要となる．紙と電子の共通の特徴として，文書の外観が挙げられる．この文書の外観を撮影した画像をキーとして対応する文書を特定する技術に，文書画像検索がある．その中でも，ユーザビリティの高いモバイル端末で動作する文書画像検索法として，Mobile Retriever [1] や Hot Paper [2]，Jorge Moraleda らの手法 [3] が提案されている．しかし，これらの手法にはサービスを実用化する上で重要なスケーラビリティやリアルタイム性が十分でないという問題がある．

これらの問題を解決する手法として，竹田らの手法 [4] がある．これは，Locally Likely Arrangement Hashing (LLAH) [5] を改良したものであり，1,000 万ページのデータベースから精度 98.9% ，処理時間 14.9ms で検索可能な手法である．以後，竹田らの手法を従来手法とする．また，これを応用し，スマートフォンでリアルタイム検索可能なシステムも開発されている [6]．このように，従来手法はある程度のスケーラビリティとリアルタイム性を持つ．一方，実用化には一般的な図書館規模での検索が必要であると考えており，これには 1 億ページの検索規模が求められる．これは，従来手法で賄いきれる規模ではなく，より一層のメモリ削減が必要となる．しかし，従来手法のメモリ削減法を用いてさらにメモリ使用量を削減すれば，検索精度に影響が生じる．従って，より安定した検索を実現する効率の良いメモリ削減法が求められる．

本稿では，上記の問題を解決するための手法を 2 つ提案する．第一に，検索に有効な特徴量を優先的にデータベースに登録する，特徴量のサンプリングである．検索での有効性とは，検索時に生じる外乱による特徴量の変動に頑健であるということである．第二に，特徴量の柔軟な照合を実現するための多重探索法である．1,000 万ページのデータベースを用いた実験から，従来手法と比較して，検索精度を維持したまま 70% のメモリ削減を実現できることが分かった．また，1 億ページのデータベースから，必要メモリ量 236GB，精度 98.7% ，処理時間 26.8ms で検索でき，高いスケーラビリティを持つことを確認した．

2. 関連研究

媒体によらないサービスを実現するインターフェースには，サービスを提示するディスプレイとデジタルカメラを持ち，携帯性に優れたスマートフォンが望ましいと考えられる．従って，スマートフォンという高度な処理能力を持たないデバイスでも動作する文書画像検索法が求められる．ここでは，スマートフォンのような携帯端末で動作する文書画像検索の関連研究について 3 つ紹介する．

“Mobile Retriever” [1] は，“token pair” と “token triplet” という，文書中のテキストに基づく 2 つの特徴量を用いる手法である．token pair は複数の単語の shape code で定義されるものであり，画像の小さな領域から正解候補の文書画像を検索

するために用いられる．token triplet は 3 つの単語とそのオリエンテーションで構成されるものであり，false positive を除去し，検索結果を上記候補から求めるために用いられる．Mobile Retriever はある程度大規模なデータベースから高い精度で検索できることが確認されている．しかし，単語の認識に OCR を用いているため，1 クエリあたり 4 秒と長い処理時間を必要とする問題がある．

“HotPaper” [2] は，Brick Wall Coding Features (BWC) と呼ばれる，複数の単語のバウンディングボックスを表現する局所特徴量に基づく手法である．バウンディングボックスのアスペクト比を用いて局所的な単語クラスタリングを行い，クラスタ毎に複数の特徴量を計算する．この特徴量はスケールに変化に不変であり，軽微な射影変換に対応できる．1 クエリあたり 250ms とある程度のリアルタイム性を備えているが，検索精度は 60% 以下であり，データベースのサイズは 5,000 ページ以下と非常に小さいものである．

Jorge Moraleda らの手法 [3] も，単語のバウンディングボックスに基づく特徴量に基づく手法である．あるバウンディングボックスの下方へジグザグにバウンディングボックスを選択していき，その中心を線で結ぶ．その線と垂直のなす角の角度と，バウンディングボックスの長さを組み合わせて特徴量とする．この手法では，50 万ページと比較的大規模なデータベースを構築しているが，特徴量の記述に 100ms を要し，検索精度も 70%程度と低いものである．

以上に示すように，どの手法にもスケーラビリティや検索精度，処理時間に問題がある．これらに対し，我々は実時間検索が可能な大規模文書画像検索法を提案している [4]．これは，LLAH に対して必要メモリ量の削減と特徴量の識別性・安定性向上を行ったもので，1,000 万ページのデータベースから精度 98.9% ，処理時間 14.9ms で検索可能な手法である．また，スマートフォンからでも 1 クエリあたり 100ms で検索できるシステムも開発されている．一方，この手法で用いているメモリ削減法では，さらなる大規模化は困難であるという問題がある．詳細については，4.3 で述べる．

3. LLAH を用いた文書画像検索

従来手法と提案手法の基盤となる LLAH を用いた文書画像検索法について説明する．

3.1 特徴点抽出処理

まず，入力画像を適応二値化する．次に，二値画像をガウシアンフィルタでぼかし，再度二値化を行うと，単語ごとに連結された画像が得られる．最後に連結成分の重心を計算し，それらを特徴点とする．

3.2 特徴量計算

特徴量とは，データベースから対応点を検索するためのキーとなるものである．特徴量には，局所領域において射影不変量に近似できるアフィン不変量を利用する．これは，検索質問画像の撮影時に生じる射影歪みに対して頑健にするためである．

アフィン不変量は，同一平面上の 4 点を頂点とする 2 つの三角形の面積比である．単一のアフィン不変量では識別性が低いため，複数のアフィン不変量を組み合わせたものを特徴量とする．具体的には，近傍 $m(>4)$ 点から 4 点を選ぶ全ての組合せから計算する．また，識別性をより高めるため，面積比特徴量を付加する．これは，単語領域の面積比に着目した射影不変な特徴量であり，m 次元の離散化されたベクトルが得られる．これを離散化されたアフィン不変量の列 $(r_{(0)}, r_{(1)}, \cdots, r_{({}_mC_4-1)})$ に追加する．従って，特徴量の次元数は $({}_mC_4+m)$ となる．

射影歪みを受けた場合，近傍 m 点として異なるものが得られることがある．しかし，近傍 n 点のうち $m(\leqq n)$ 点までは，共通のものが得られる可能性が高い．そこで，n 点から得られるすべての m 点の組合せから特徴量を計算する．従って，1 特徴点あたり ${}_nC_m$ 個の特徴量が計算されることになる．

3.3 登録処理

特徴量をハッシュ表のインデックス H_{index} に変換し，それに基づいて特徴点をハッシュに登録する．以下で示すハッシュ関数を用いて H_{index} を求める．

$$\left(\sum_{i=0}^{{}_mC_4+m-1} r_{(i)}d^i\right) = QH_{\mathrm{size}} + H_{\mathrm{index}} \tag{1}$$

ここで d は離散化レベル数，H_{size} はハッシュ表のサイズを表す．ハッシュ表には文書 ID，特徴点 ID，商 Q を 1 組として保存する．また，衝突が生じた場合は，リスト形式で追加する．リストの長さには制限値を設け，その制限値を超えた場合には，リスト全てをインデックスごと削除する．以降，削除されたインデックスは使用しないこととする．

3.4 検索処理

登録処理と同様に各特徴点に対し，H_{index} と商 Q を求める．H_{index} を用いてハッシュを参照し，リストを取得する．リストの各項目について商 Q が一致するかを調べ，一致した場合にそのリストの文書 ID に投票する．最後に，最大得票数を得た文書を検索結果とする．

4. 従来手法

4.1 特徴量の識別性・安定性向上

データベースの大規模化に伴い，類似した特徴点の配置を持つ文書が登録される可能性が高くなる．これにより誤投票が増加し，検索精度が低下する．これを防ぐため，m の値を大きくし特徴量の次元数を増加させることで，識別性を向上させることを考える．しかし，次元数が増加すれば，検索時に特徴量のマッチングが得られなくなる可能性が高くなる．これは，特徴点の位置が少なからず変動し，アフィン不変量に誤差が生じるためである．検索では，特徴量の各次元がすべて一致する必要があるため，特徴量の安定性を向上させる必要がある．

この問題を解決するため，次元数を増加させたうえで冗長な次元を除去することで，識別性と安定性の向上を両立させる．LLAH のアフィン不変量の算出には 2 つの三角形を用いているが，この三角形が他の不変量と重複している場合がある．この場合，双方の不変量には高い相関があり，冗長であると考えられる．そこで，不変量が同一の三角形を共有しないように不変量を計算することで，特徴量の冗長性を解消する．これにより，

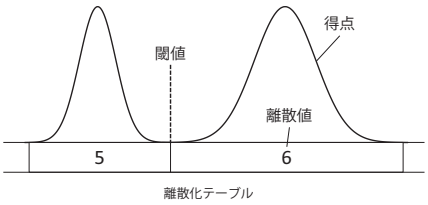


図 1　離散化の閾値からの距離による得点

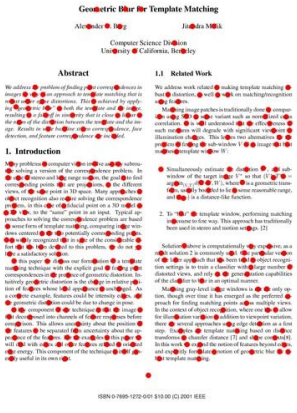


図 2　特徴量の分布

特徴量はアフィン不変量ベクトル $({}_mC_4 - 1)/2$ 次元，面積比特徴量 m 次元の計 $({}_mC_4 - 1)/2 + m$ 次元となる．

4.2　特徴点のサンプリング

LLAH では，検索に投票処理を用いているため，データベースにすべての特徴点を保存していなくとも検索可能である．そこで，データベースに登録する特徴点をサンプリングすることにより，必要メモリ量の削減を図る．

従来手法では，以下の手順でサンプリングする特徴点を決定する．まず，ある特徴点 s の抽出元となる単語領域の面積を計算する．次に，特徴点 s の近傍 m 点の単語領域と面積を比較する．特徴点 s の単語領域の面積が最小となったとき，これをデータベースに登録する．このような極小領域となる単語は文書全体に分布しているため，文書中の位置に偏りのないサンプリングが実現できる．また，抽出された特徴点は近傍点との距離が小さくなるため，射影歪みの影響を受けにくい特徴量が得られる．従来手法では，上記手法で得られた特徴点に加え，1 文書あたりの特徴点数が 200 点になるよう，その近傍 k 点もハッシュに登録する．k の値は文書ごとに設定する．特徴点数が 200 点に満たない場合はサンプリングしないこととする．

4.3　従来手法の問題点

従来手法の問題点として，さらなる大規模化への対応が困難であることがあげられる．仮に 1 億ページのデータベースを作成した場合，メモリ使用量が約 700GB と膨大になってしまう．従って，今以上のメモリ削減が必要であるが，特徴点のサンプリングをこれ以上行うと，検索精度に影響が出る．また，文書上で特徴点が登録されていない領域が多くなり，部分検索への耐性が低下する．以上の理由から，これ以上の特徴点のサンプリングは望ましくない．

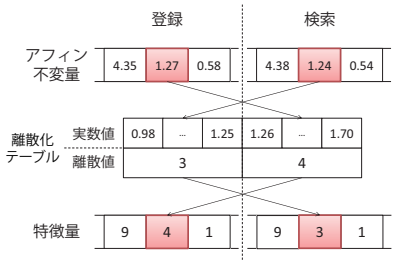


図 3　特徴量の不一致

5.　提 案 手 法

上記の問題を解決するため，2 つの手法を提案する．第一に，特徴量のサンプリングによるメモリ削減法である．第二に，多重探索による検索の安定性向上法である．

5.1　特徴量のサンプリング

特徴量のサンプリングでは，検索に有効な特徴量を優先してハッシュ表に登録することで，必要メモリ量の削減と安定した検索を実現する．ここで，本研究における検索に有効な特徴量とは，文書撮影時の外乱によって生じるアフィン不変量の変動に強く，マッチングが得やすいものである．つまり，ある程度の変動が生じたとしても，同一の離散値に変換される不変量で構成される特徴量のことを指す．

検索時に生じる誤差を，特徴量の登録時に把握することは困難である．しかし，変動が生じた場合に，登録時と異なる離散値に変換されやすいかどうかは，離散化の閾値との距離からある程度予想できる．つまり，登録時に離散化の閾値に近い不変量は，変動によって異なる離散値に変換されやすいということである．従って，閾値から遠い不変量で構成された特徴量がより高い安定性を持つことになる．提案手法では，離散化の閾値から離れた不変量で構成される特徴量を検索に有効なものと捉え，それらを優先してデータベースに登録する．これにより，必要メモリ量の削減と検索精度の低下の抑制との両立を図る．

特徴量のサンプリング法について述べる．まず，離散化テーブルの各離散値領域に対して正規分布を適用し，領域の端に行くほど低く，中心に行くほど高くなるように点数を割り当てる．点数の割り当て例を図 1 で示す．このように，離散化の閾値からの距離に応じた点数を用いることで，特徴量の有効性を判定する．次に，不変量を離散化する際に，対応する離散値に変換するとともに，離散化領域内の位置に応じて割り当てられた点数を付与する．そして，各特徴量の中で不変量の点数を比較し，最小得点をその特徴量の有効性の点数とする．これは，1 次元でも変動に弱いものがあれば，検索に失敗する可能性が高いためである．データベースへの登録時には，得点の高い特徴量から順にデータベースに登録する．登録する特徴量数は 1 文書あたり N 個になるように設定する．全特徴量数が N 個に満たない場合は，全ての特徴量をハッシュに登録する．

文書全体から特徴量を選択すると，文書画像中における特徴量の位置の分布に偏りができてしまい，検索できる範囲とでき

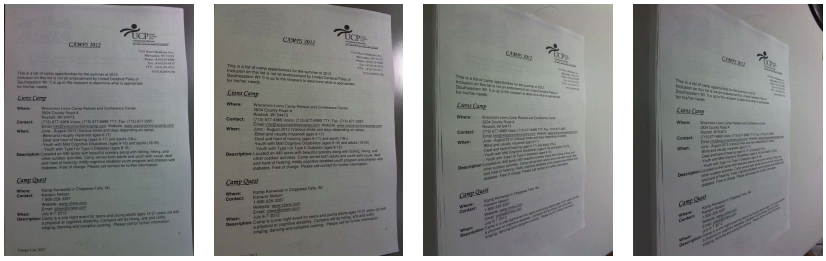
図 4　登録文書画像

表 1　手法の性能比較

登録文書画像数	手法	検索精度 [%]				必要メモリ量 [GB]	処理時間 [ms]
		90°	75°	60°	45°		
1,000 万	従来手法	98.9	98.3	95.3	76.1	99.3	14.9
	提案手法 1	98.9	99.0	95.1	72.9	55.0	13.0
	提案手法 2	99.1	98.7	95.1	76.6		29.9
1 億	提案手法 1	98.7	98.6	92.4	60.7	236	26.8
	提案手法 2	98.4	97.9	90.3	58.0		57.1

(a) 単語の少ない文書　(b) 画像の多い文書

図 5　除外された文書画像



(a) 90°　(b) 75°　(c) 60°　(d) 45°

図 6　検索質問画像

ない範囲が生じてしまう．そこで，文書画像を複数の領域に分割し，各領域に含まれる特徴点数に応じて登録する特徴量数を決定する．登録された特徴量の文書上での位置を図 2 に示す．このように，単語が含まれる領域に特徴量が分散しており，部分撮影された検索質問画像も検索可能となる．

5.2 多重探索

変動に強い特徴量を登録しても，検索時に生じる全ての変動に対応できるわけではない．少なからず，登録時と異なる離散値に変換されてしまう不変量も存在する．このような場合，離散化前の不変量は離散化の閾値から近い位置にあると考えられる．そこで，検索時に閾値に近い不変量に対しては，閾値前後の 2 つの離散値を与えることを考える．これにより，ある次元の離散値が異なる 2 つの特徴量が生成されることになり，どちらか一方が登録時と同じ特徴量になる可能性が高くなる．このような処理を多重探索と呼ぶ．

提案手法では，複数の次元で上記のような 2 通りの離散値が得られた場合，離散値の組み合わせ全てを用いて特徴量を生成する．従って，生成される特徴量数は，2 の累乗で増加していくことになる．これにより，登録時の特徴量と同一のものが得られやすくなるが，一方で誤投票が増加する可能性もある．また，特徴量の生成やハッシュの参照回数の増加により，処理時間が増加すると考えられる．

6. 実　　験

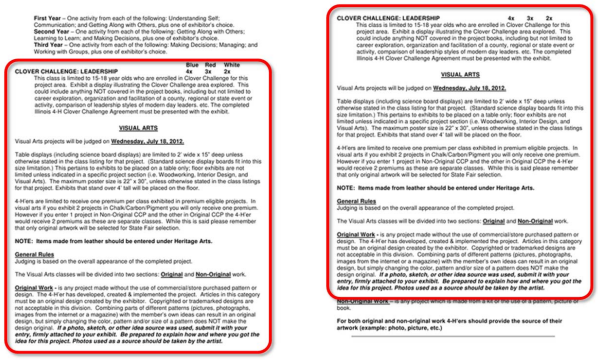
6.1 実験 1：提案手法の性能とスケーラビリティの検証

6.1.1 実験概要

本稿で述べた特徴量のサンプリングと多重探索の効果を検証するため，メモリ使用量と検索精度及び処理時間の比較実験を行った．ここで，処理時間は検索質問画像 1 枚当たりの平均であり，特徴点抽出にかかる時間は含まれない．比較する手法として，従来手法，特徴量のサンプリングのみを行った提案手法 1，そして特徴量のサンプリング及び多重探索を行った提案手法 2 の 3 つを用いた．

データベースの作成に用いる文書画像として 1 億ページ用意した．これらは，インターネットから収集した PDF ファイルを 200dpi で画像に変換したものである．登録文書の重複を避けるため，各 PDF の MD5 ハッシュ値を用いてファイルの同一性を確認し，重複する PDF を除外した．登録文書画像の例を図 4 に示す．本実験では，1,000 万ページを用いたデータベース A と，1 億ページと用いたデータベース B を作成する．データベース B はデータベース A を包含している．まず，A に対する各手法の性能を比較することで，提案手法の有効性を検証する．次に，B に対する提案手法 1 と提案手法 2 の性能を確認することで，手法のスケーラビリティを検証する．

検索質問用の文書画像として，データベース A に用いた文書画像の中から以下の手順で 1,000 ページを抽出した．まず，1,000 万ページの中からランダムに 2,000 ページを抽出する．次に，2,000 ページの中から，特徴点数が 100 以下のものと，グレースケール画像ヒストグラムにおいて画素値が 250 以上の割合が 70%以下のものを除外する．除外された画像例を図 5 に示す．このような単語数の極端に少ない文書画像と，文字以外が大半を占める画像を取り除くことで，純粋な文書画像検索の精度を検証することができる．最後に，残ったものの中からランダムで 1,000 枚抽出する．これまでの LLAH を用いた実験では，科学技術論文のみを検索質問画像としてきたため，ある特定の条件下での性能しか計測できていなかった．これに対し，ランダムに検索質問を選択することで，文書画像検索の精度に一般性を持たせることができる．抽出した 1,000 枚を印刷し，文書全体が写るように紙面に対して仰角 90°, 75°, 60°, 45° から撮影したものを検索質問画像とした．検索質問画像



(a) 正解画像　　(b) 検索結果

図 7　類似画像への対応

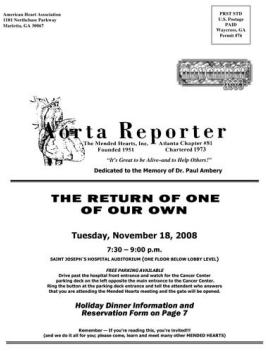


図 8　特徴点の抽出失敗例

図 9　リスト形式の文書

の解像度は，$2,592 \times 1,944$ である．撮影には，Logicool HD Pro Webcam C910 を用いた．検索質問画像の例を図 6 に示す．ハッシュ表のサイズは $2^{32} - 1$ である．計算機は，CPU が AMD Opteron 2.9GHz，メモリが 512GB のものである．パラメータは，$n = 8$，$m = 7$，$N = 400$ とした．実験結果を表 1 に示す．

6.1.2　必要メモリ量

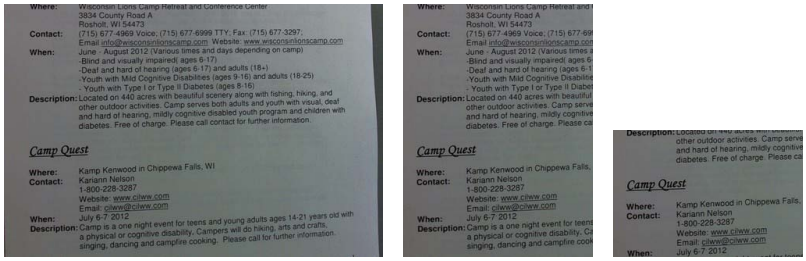
1,000 万ページデータベースにおいて，従来手法の必要メモリ量は 99.3GB，提案手法 1 では 55.0GB となった．提案手法 2 は提案手法 1 と同じデータベースを用いているため，必要メモリ量の値は同じである．必要メモリ量には，ハッシュテーブルを確保するために必要な 32GB が含まれている．従来手法において，登録特徴量数が 104 億だったのに対し，提案手法では 34 億だったとこから，約 70% のメモリ削除を実現したことになる．1 億ページデータベースにおける必要メモリ量は 236GB となり，登録特徴量数は 328 億であった．

6.1.3　検 索 精 度

従来手法と提案手法 1 を比較してみると，1,000 万ページのデータベースにおいて，90°〜60° の検索質問画像に対してほぼ同程度の検索精度となっていることがわかる．ここで，90° の検索質問の 1 クエリあたりの平均得票数を見てみると，従来手法で 206，提案手法 1 で 100 であった．データベース中の特徴量数が 70°% 削減されたことを考えると，各手法の登録特徴量数に対する得票数の割合は提案手法 1 の方が大きい．従って，特徴量のサンプリングによって，検索に有効な特徴量が選択できたといえる．次に，提案手法 1 と提案手法 2 を比較すると，特に 45° の検索質問に対して提案手法 2 が良好な結果を示している．仰角 45° から撮影すると射影変換が強くかかり，特徴量が大きく変動してしまう．この影響を，多重探索によって修正できたことの表れであると考えられる．

次に，1 億ページのデータベースに対する検索精度を見ていくと，提案手法 1 において 90° の検索質問画像に対して 98.7% と高い精度を実現できていることが分かった．一方，提案手法 2 の検索精度は提案手法 1 よりも低くなった．これは，多重探索による誤投票の影響であると考えられる．多重探索では，生成された特徴量のいずれかが正しい特徴量とマッチングし，それ以外の特徴量にはマッチングするものが存在しないことが望ましい．しかし，生成特徴量の中には正解以外の特徴量にマッチングするものも少なからず存在する．1,000 万ページデータベースでは，ハッシュテーブルのインデックス占有率は 41% と少なく，平均リスト長も 1.9 と短いため，誤投票が生じる可能性は低い．ここで，平均リスト長とは，ハッシュ表のリスト長が 0 でないインデックスのリスト長の平均である．一方，1 億ページのデータベースでは占有率が 95% 以上であり，平均リスト長が 8 と長いため，誤ったマッチングが生じる可能性が高くなっている．この誤投票によって得票数の順位が入れ替わり，検索に失敗したものが発生したと考えられる．このことから，多重探索の有効性は，データベースに依存することが分かった．

最後に，検索に失敗した画像について考察する．誤検出された結果を見ると，失敗の原因が大きく 3 種類に分類できることが分かった．第一に，類似文書への対応である．この例を図 7 に示す．図 7 の枠で囲んだ部分のように，文書の一部分が重複しているものや，レイアウトが全く同じで記述されている単語が異なるものに対応づいたものがあった．このような場合，文書のレイアウトに依存する LLAH の特徴量では完全に識別することは困難であると考えられる．第二に，特徴点抽出の失敗による影響である．図 8 に示すように，極端に文字が大きな文書や，文書中で文字の大きさが変化するものからは安定して単語領域を取り出すことが難しい．特徴点が安定して抽出できなければ特徴量も安定しないため，検索に失敗したものと考えられる．これに対処するため，より安定した単語抽出法を考案する必要がある．第三に，単語が規則的に並んでいるリスト形式の文書への対応付きである．図 9 のように，単語の配置が規則的であったとき，同じ文書から同じ特徴量が大量に生成される．ある検索質問画像に別の文書と同じ特徴量が存在したとき，誤投票になるが，通常であれば誤差で済む．ところが，同じ特徴量が同じ文書 ID で大量に登録されている場合，その文書に投票が集中してしまい，そちらに引っ張られて検索失敗となってしまう．このような例が検索失敗の 50% を占めていた．従って，同一文書に同一特徴量が大量に発生した場合に，その代表特徴量のみを保存するといった改善が必要になると考えられる．



(a) 2 分の 1　(b) 4 分の 1　(c) 8 分の 1

図 10　部分撮影の検索質問画像

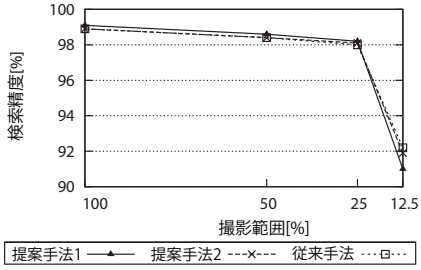


図 11　部分撮影での検索精度

6.1.4 処 理 時 間

1,000 万ページデータベースで従来手法と比較して，提案手法 1 の方が処理時間が短くなった(注1)これは，ハッシュ表の平均リスト長の違いによるものである．平均リスト長は，従来手法では 3.2 であるのに対し，提案手法 1 では 1.9 であった．この違いは，ハッシュに保存されている特徴量数の違いによるものである．平均リスト長が長いほど，リストを参照する回数が増加するため，処理時間が増加する．提案手法 2 の処理時間が最も長くなった原因は，多重探索で新たな特徴量を生成することにより，ハッシュを参照する回数が増加するためである．多重探索によって 1 クエリあたりに平均して生成される特徴量数は約 1 万であり，多重探索なしの場合と比較して約 3 倍となっている．この違いにより，処理時間が増加したと考えられる．しかし，それでも 30ms 以下の処理時間となっており，良好な結果を残していると考えられる．

1 億ページのデータベースに対しては，提案手法 1 で 1 検索質問あたり 26.8ms と実時間検索可能な処理時間を実現できた．また，提案手法 2 では 57.1ms と提案手法 1 の 2.1 倍となった．

6.2 部分撮影への耐性

特徴量のサンプリングによって大量の特徴量を削除した場合，全ての特徴点がデータベースに登録されるわけではない．これによって，部分的に撮影された検索質問への耐性がなくなってしまうことが考えられる．そこで，文書を部分的に撮影した検索質問画像を用いて実験を行い，部分撮影への耐性を検証する．

検索質問画像は，実験 1 で用いた 1000 枚の画像から一部分を切り抜くことで，疑似的な部分撮影クエリを作成した．部分撮影の大きさは，文書全体の 2 分の 1，4 分の 1，8 分の 1 の 3 パターン用意した．切り抜く際には，文書が映っている範囲を選ぶようにした．検索質問画像の例を図 10 に示す．その他の条件は実験 1 と同じである．

検索精度を図 11 に示す．従来手法と比較して，提案手法 1 は同程度の検索精度を実現できていることがわかる．従って，提案手法 1 の特徴量のサンプリングは，大量に特徴量を削減しても部分撮影に耐えうる手法であることが示されたといえる．

撮影範囲が狭くなるにつれて，検索精度が低下した．これは，検索質問画像に含まれる特徴点数が少なくなり，検索必要な特徴点数が得られなかったことが原因であると考えられる．しかし，文書全体の 8 分の 1 の大きさであっても 92%以上の精度を実現していることから，部分撮影への耐性は十分保持していると考えられる．

(注1)：[4] よりも処理時間が短縮されているのは，投票テーブルの初期化処理の改善によるものである．

7. ま と め

本稿では，必要メモリ量の削減と検索の安定性を両立させるため，特徴量のサンプリングと多重探索法を提案した．実験結果から，特徴量のサンプリングを用いた手法で，検索精度を維持したまま必要メモリ量を約 70% 削減できることが分かった．また，1 億ページのデータベースから，検索精度 98.7%，処理時間 26.8ms で検索できることを確認した．さらに，部分撮影への耐性実験から，文書全体の 4 分の 1 の大きさの検索質問であっても 98% 以上の精度で検索できることも分かった．

今後の課題としては，さらなる大規模化のためにより効率の良いメモリ削減法を考案することがあげられる．また，より安定した特徴点抽出法を考案する必要もある．

謝　　辞

本研究の一部は，JST CREST および日本学術振興会科学研究費補助金基盤研究（B）(22300062)，挑戦的萌芽研究 (21650026) の補助による．

文　　献

[1] X. Liu and D. Doermann, “Mobile retriever: access to digital documents from their physical source,” Int. J. Doc. Anal. Recognit., vol.11, pp.19–27, Sept. 2008.

[2] B. Erol, E. Antúnez, and J.J. Hull, “Hotpaper: multimedia interaction with paper using mobile phones,” Proceeding of the 16th ACM international conference on Multimedia, pp.399–408, 2008.

[3] J. Moraleda and J.J. Hull, “Toward massive scalability in image matching,” ICPR, pp.3424–3427, 2010.

[4] K. Takeda, K. Kise, and M. Iwamura, “Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah,” 2011 International Conference on Document Analysis and Recognition, pp.1054–1058, Sept. 2011.

[5] T. Nakai, K. Kise, and M. Iwamura, “Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval,” Lecture Notes in Computer Science (7th International Workshop DAS2006), vol.3872, pp.541–552, feb 2006.

[6] K. Takeda, K. Kise, and M. Iwamura, “Real-time document image retrieval on a smartphone,” Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS2012), pp.225–229, March 2012.