

類似検索における複数特徴量間の重みを考慮した データベースアクセスの最適化

谷口 展郎[†] 高橋 達也[‡] 山室 雅司[†]

[†]NTT情報通信研究所 [‡]岩手大学

画像や音声のデータから特徴量を抽出し、その類似度に基づいて検索を行う「特徴量類似検索」においては、一般に複数の特徴量に重みづけをして検索を行うことが多い。こうした場合、検索の各特徴量空間への振り分けかたが、検索の正確さや効率に大きな影響を与える。我々は与えられた重みに基づいてこの振り分けを最適化する2つの手法を提案し、うち一つについては実験により有効性を確認したことまでを先に報告した。本稿では、残るもう一つの手法、および2手法の組み合わせた場合の実験結果を中心に報告する。

Access Optimization for Content-based Retrieval using Weighted Multiple Features

Noburou Taniguchi[†] Tatsuya Takahashi[‡] Masashi Yamamuro[†]

[†]NTT Information and Communication Systems Laboratories

[‡]Iwate University

Content-based retrieval is a promising technique for retrieval of information such as image, sound and video. For this type of retrieval, it is usual that multiple features are used for similarity evaluation. In such case, accuracy and efficiency of retrieval much depends on how a multi-feature query is translated into search of each individual feature. We have previously reported that we proposed two methods to optimize translation and showed effectiveness of one of the methods by experiment. In this paper, we show additional empirical observation of the other method, and the combination of the methods.

1. はじめに

デジタルカメラの普及等により、画像のデジタル化も急速に進展しており、大量のデジタル画像の管理への要求も高くなっている。我々はこれまで、画像から自動抽出した特徴量をもとに検索を行うシステム ExSight+HyperMatch の研究を通じて、この問題に取り組んできた [2, 3]。

特徴量検索 (FBR: Feature-Based Retrieval) は、検索キーとして与えられた画像と、検索対象としてデータベース中に貯えられた画像の、それぞれから抽出された特徴量を比較し、類似度の高いものを返すという手法である。画像の FBR で

は、一般に、色、模様、構造、形状などを表す複数の特徴量を組み合わせ、それぞれに重みづけして検索を行うことでユーザーの多様な検索意図を反映できる仕組みをとるものが多い [9, 10]。

FBR では、特徴量を多次元空間中のベクトルとして扱うことが一般的で、高速化には R-tree [11] を代表とする多次元木構造ベースのアクセス手法を用いることが多い。検索エンジン HyperMatch でも、個々の特徴量ごとに張られた木構造インデックスを利用して重みつき複数特徴量検索を行っているが、この場合、検索の各特徴量への振り分け方に十分な考慮を払わないと、最終的な検

索結果の正確さが低下したり、余分な処理のために検索効率が低下したりする。

従来、我々のシステムでは、検索の各特徴量への振り分けかたが固定的であり、これが上述の検索の正確さや効率を向上を図るうえでの課題であった。そこで我々は、ユーザーから与えられた各特徴量の重みに応じてパラメータを変化させて検索を振り分ける2つの手法を提案し、そのうち1つについては、実験により、最終検索結果の正確さを低下させることなく性能が向上することを確認した(既報[1])。

本報告では、残されたもう一つの手法、およびこれら2手法の組み合わせについて行った実験の結果を中心に述べる。

2. 重みを考慮したアクセス最適化

2.1. 重みつき複数特徴量類似検索

まず、本報告が対象とする検索および提案手法について述べる。既報と重なる部分も多いが、よろしくご承知いただきたい。

前述のように、FBRでは特徴量を多次元空間中のベクトル(特徴ベクトル)として扱うことが多い。この特徴量空間において、例えば「この画像に似た画像上位10件を探せ」という問い合わせが、「このベクトルに近いベクトル上位10件を探せ」という検索に置換されて実行される。

多次元木構造インデックスは、上記のような検索を高速化するために用いられる。まず、対象とする空間中で、距離が近いベクトルを階層的にグループ化して木構造を構築する。検索時には、各階層においてキーベクトルの近くにあるグループを辿ることで、上位にくる可能性が高いベクトルに限定して距離計算を実行する。こうして、コストの大きい距離計算の回数を削減し、検索効率を向上させることができる。

検索が単一の特徴量に限られていれば、これで特に問題はない。しかし実際の画像検索においては、複数の特徴量に、その特徴量をどれだけ重視するかという重みを与え、これらを組み合わせて利用することが多い。こうした検索においては、複数の特徴量の重みつき距離 d ,

$$d = \sum w_i d_i \quad (\text{式1})$$

w_i : i 番目の特徴量の重み

d_i : i 番目の特徴量での距離

を用いて総合的な類似度を評価するのが一般的である。この場合、当然これらの重みづけは動的に

変更できることが望ましい。

複数特徴量を用いる検索において木構造インデックスを利用する場合、(A) 複数の特徴量空間から単一の多次元空間を合成し、その合成空間に対して木構造インデックスを構築する (B) 個々の特徴量空間ごとに木構造インデックスを構築するという2通りの方法が考えられる。

(A) を採用した場合、特徴量間の関係は木構造インデックスが構築されたときに固定されるため、特徴量の重みづけを動的に変更することは基本的に不可能である。いろいろな重みの組み合わせに対してそれぞれ木構造インデックスを構築しておき、検索時に切り替えて使うという方法も考えられなくはないが、組み合わせの数から考えて非現実的であろう。

一方、(B) を採用した場合、まず個々の特徴量において検索を行ったうえで、これらの検索結果を統合し、最終検索結果を求めるという過程を経ることになる。ここで重要になってくるのが「総合結果の上位 r 件を正確に得るためには、個々の特徴量からはどれだけ検索結果が必要か」という問題である。例えば「この画像に似た画像上位10件を探せ」という検索を実行する場合、「個々の特徴量から上位10件ずつ集めて統合すればよい」というものではない。なぜなら、各特徴量において全て11位の類似度を与える画像が存在して、それが総合的な類似度において10位以内になる可能性も考えられるからである。

実際には、非常に限られた場合(例えば「全ての特徴量空間でデータが一様分布」など)を除いて、ユーザー要求件数 r から各特徴量の検索回答数 r_i を正確に予測することは基本的に不可能である。つまり、複数特徴量を用いた検索で (B) の方法をとった場合、総合上位 r 件の「100%」の正確さは保証されない。もちろん、少しでも正確な総合結果を得るためには、例えば r_i を大きく取ればよいが、これにより検索コストは増大する。

つまり、複数特徴量を用いた検索においては、

- ・総合結果の正確さ
- ・総合結果の検索コスト

の両方が高い水準でバランスするよう、いかに個々の特徴量検索を最適化するかが重要になる。

2.2. 重みを考慮したアクセス最適化

我々は以下の2つのアイデアを基本に、重みつき複数特徴量類似検索の最適化手法を考えた。

- ・個々の特徴量の検索結果は、あくまで総合結果の候補にすぎない

- ・重みが大きい特徴量ほど、その検索結果が総合結果に与える影響は大きい

個々の特徴量検索では、必ずしも上位から順に正確に並べて返す必要はない。例えば、100%正確な上位 m 件を返すより、ある程度正確な上位 $m+n$ 件を返すほうがコストが小さければ、後者の手段を採用したほうがよいと考えられる。

さらに、こうした個々の特徴量検索の「回答数」「正確さ」は、他の特徴量とのバランスの上で決定されるべきものである。なぜなら、これら複数の特徴量の検索から得られた集合が、「ある特徴量からの検索結果中に m 件、別の特徴量からの検索結果中に n 件、…」というように、トータルで総合順位上位 r 件を含む総合結果の候補群を得ることがこの処理の本質だからである。

一方、式1から、検索時に与えられた重みが大きい特徴量ほど、その検索結果が総合結果に与える影響は大きいことは明らかである。したがって、各特徴量の重みを、この「正確さ」「回答数」のバランスを決める最も主要なファクターと考えるのが自然であると考えた。

以上、重みつき複数特徴量類似検索の最適化に対して、我々は、個々の特徴量検索における「回答数」「正確さ」を、その特徴量の重みを考慮して決定するというアプローチを取ることとした。

具体的には、「回答数」「正確さ」それぞれを決定する方法として、以下の2つを提案した。

◆回答数の決定方法

個々の特徴量検索からの回答数 r_i は、

$$r_i = m_i r$$

とした。ここで余裕係数 m_i は、該特徴量の重み w_i をパラメータとして、

$$m_i = f_m(w_i)$$

で表される関数により決定される値である。

◆正確さの決定方法

これについては、多次元木構造インデックスにおける近似近件検索アルゴリズム [8] を利用する。このアルゴリズムの詳細については参考文献を参照していただくことにし、ここでは概要を述べるにとどめる。

ある特徴量空間においてキー Q に近いベクトル上位 r_i 番目までを求める検索を行うものとする。多次元木構造インデックスを利用してこの検索を行う場合、2つのリストが使われる。一つは今まで辿った部分木中で、まだアクセスしていない内部ノード（下位に子ノードを持つノード）を

Q からの距離順に並べたもの L_{int} 。もう一つは、既にアクセスした葉ノード（実ベクトルを表すノード）を Q からの距離順に上位 r_i 件まで並べたものである L_{leaf} である。

今、 Q から L_{int} 中の最上位ノード P_{itop} までの距離 d_{imin} 、 L_{leaf} 中の最下位 (r_i 位) ノード P_{ilast} までの距離を d_{imax} とする (図1)。 d_{imin} が d_{imax} より小さければ、ノード P_{itop} にアクセスし、その子ノード群と Q との距離を計算して、 L_{int} または L_{leaf} を更新する。もし、 d_{imin} が d_{imax} より大きければ、そこで検索を終了し、 L_{leaf} を検索結果として返す。

近似検索アルゴリズムでは、エラー率 $e (> 0)$ を導入して、上記の d_{imax} を $d_{imax} / (1 + e)$ に置き換える。これにより閾値が小さくなるので、 P_{itop} へのアクセスの可能性が低下し、結果として検索が早く終わる。このとき、検索結果の r_i 位のベクトルを P 、真の r_i 位のベクトルを P' とすると、 $QP = (1 + e) QP'$ であることがわかっている。

我々は、近似係数 a_i (上記アルゴリズムでは、 $1 / (1 + e)$ に相当) により、個々の特徴量検索の正確さを定義する。 a_i は、該特徴量の重み w_i をパラメータとして、

$$a_i = f_a(w_i)$$

で表される関数により決定される値である。

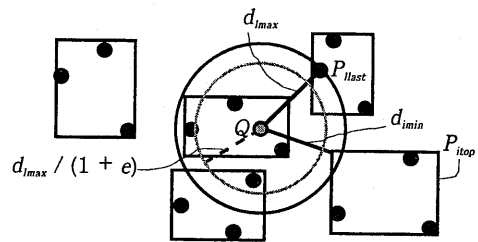


図1 検索アルゴリズム

3. 検索システム ExSight + HyperMatch

本報告で実験を行った画像検索システム ExSight+HyperMatch の構成を図2に示す。検索に利用した特徴量は以下の8種である。

- ・色に関する特徴量 (4種類)
 - …色相, 彩度, 明度, 主観色基準
- ・形に関する特徴量 (2種類)
 - …外周形状 [2], ウェーブレット [4]
- ・幾何情報の特徴量 (2種類)
 - …面積, 重心位置

以下、本システムの検索処理の概要を示す。詳

細については、文献 [1-3] を参照されたい。

(1) 検索キーおよび検索条件の入力

検索キーは、画像として与えられたものを、

- ・ 特徴量ベクトルの組 $\{v_1, \dots, v_i\}$

に変換する。このほかに、検索条件として、

- ・ 要求回答数 r
- ・ 特徴量ごとの重み $\{w_1, \dots, w_i\}$
($0 < w_i < 1$)

が与えられる。

(2) 個々の特徴量の検索

(重み0の特徴量については検索を行わない)

個別特徴量検索部に渡されるのは、

- ・ キーベクトル v_i
- ・ 回答数 r_i
- ・ 近似係数 a_i

である。個別特徴量検索部は、キーベクトル v_i に類似した特徴量ベクトルをもつ画像を上位から r_i 件を a_i 近似で返す。

(3) 結果の統合→最終検索結果の返却

コーディネイター部は、個別特徴量検索部から返された全ての画像に対し、式1に従って重みつき距離 d を算出し、 d の小さい順に上位 r 件を最終検索結果として返す。

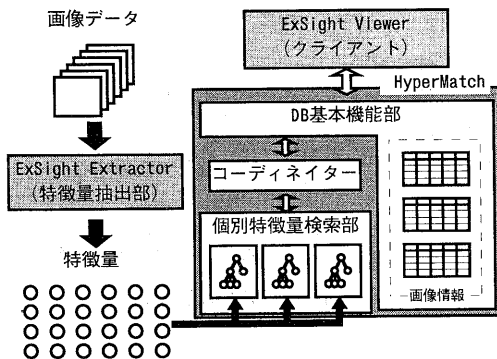


図2 システム構成

4. 実験および考察

第2節で述べた複数特徴量類似検索の最適化手法のうち、近似係数を利用する方法については、既報で述べた通り、すでに実験から有効性を確認している。今回、残された余裕係数を利用する方法の実装が完了したので、余裕係数法および両者の組み合わせ手法について、既報と同じ条件で実験を行い、有効性の評価を行った。

4.1. 実験条件

既報からの変更はないが、再掲しておく。

実験は、Sun 社の Ultra 30 ワークステーション (OS: Solaris 2.5.1) 上で行った。

データセットとして、写真1009枚から物体抽出 [2] により切り出した96481枚の子画像を用いた。この子画像それぞれから前節の8つの特徴量を抽出し、データベースに投入したのち、各特徴量についてノードサイズ 200 の VAMSplit R-tree [7] を構築した。

問い合わせは、我々のシステムで実際に画像検索に使用されたものの中から、代表的なもの12例を使用した (表1)。単純化のため、総ての問い合わせで、要求回答数 r は70に固定した。

表1 各問い合わせにおける特徴量ごとの重み

問い合わせ	1	2	3	4	5	6	7	8	9	10	11	12
色相	1	0.5	1	0.9	0.9	0.5	0.5	1	0.9	0.1	0.7	1
彩度	0.2	0.5	0.6	0.2	0.5	0.5	0.5	0.5	0.3	-	0.2	0.5
明度	0.1	-	0.1	0.4	-	-	-	0.2	0.1	-	-	-
主観色基準	1	0.8	0.5	1	0.9	-	0.5	-	0.8	0.2	0.6	1
外周	0.1	0.3	0.2	0.8	0.5	0.5	0.2	1	0.1	1	0.5	0.5
ウェーブレット	-	-	-	-	-	-	0.2	0.5	-	0.3	-	-
面積	1	1	1	-	0.5	-	-	-	0.6	0.9	-	0.9
重心位置	0.3	0.2	1	-	-	-	-	-	0.3	0.6	-	0.8

注: - は重み0で検索されないことを示す

4.2. 検索方法

個々の特徴量検索のパラメータは、以下のように設定した。

◆ 近似係数 a

- ・ 固定 (従来方式) : $a = 1$
- ・ 重み反映 : $a_i = w_i$

◆ 余裕係数 m

- ・ 固定 (従来方式) : $m = 6$
- ・ 重み反映 : $m_i = 3nw_i$

4.2. 実験結果および考察

上記方式を組み合わせた、以下の4方式

- ・ a 固定 & m 固定 (従来方式)
- ・ a 重み反映 & m 固定 (既報)
- ・ a 固定 & m 重み反映
- ・ a 重み反映 & m 重み反映

について最終検索結果の正確さおよびそのコストの評価結果を以下に示す。前2者は既報と重複するが、比較のため再掲する。

図3は、各問い合わせごとに遺漏度を示したものである。遺漏度は、 r -近件検索の正確さを評価

するための指標で、

$$o = \sum 1/R_i$$

R_i : 回答集合から漏れていた要素
の正解集合における順位

で表される。この定義から分かるように、遺漏度は小さいほどよく、遺漏度0で正解と完全に一致していることを意味する（遺漏度の詳細については既報 [1] を参照されたい）。参考として、最悪値、すなわち70件の検索結果がすべて間違っていた場合の遺漏度4.8328 ($= \sum 1/N, N = 1..70$) も図中に示した。なお、遺漏度の測定には、木構造インデックスを経由しない全件検索により別途求めた正解集合を用いた。

この図からわかるのは、大きく

- ・ m 重み反映法は遺漏度低減に有効
- ・ a 重み反映法は遺漏度に殆ど影響しない

の2点である。

余裕係数 (m) 重み反映法では、「a 固定」の場合で平均約49%、「a 重み反映」の場合で平均

約53%の遺漏度の低減を達成した。もちろん、個々の問い合わせでは効果はまちまちで、最高約90%も低減できた例（問い合わせ4）もあれば、逆に悪化した例（問い合わせ6,7）もあった。

問い合わせ6,7で遺漏度が悪化した原因は、この2つの問い合わせでは重みの最大値が1ではない（表1参照）ためであると考えられる。

余裕係数重み反映法は、「重要な特徴量は回答数を多く、重要でない特徴量からは回答数を少なく」というように、重みの相対性を利用して検索の正確さ/効率の向上を図るものである。これに対し、例えば問い合わせ6では、使用する3つの特徴量の重みが全て0.5、つまり「3つとも中くらいしか重要でない」と解釈されるように与えられているため、全ての特徴量から正確さがあまり高くない結果が返され、これが最終結果の正確さに悪影響を与えたのであろう。

この問題は、入力処理において各特徴量の重みの最大値が1になるように正規化することで簡単に解決できる。しかし、ユーザーの真の検索意

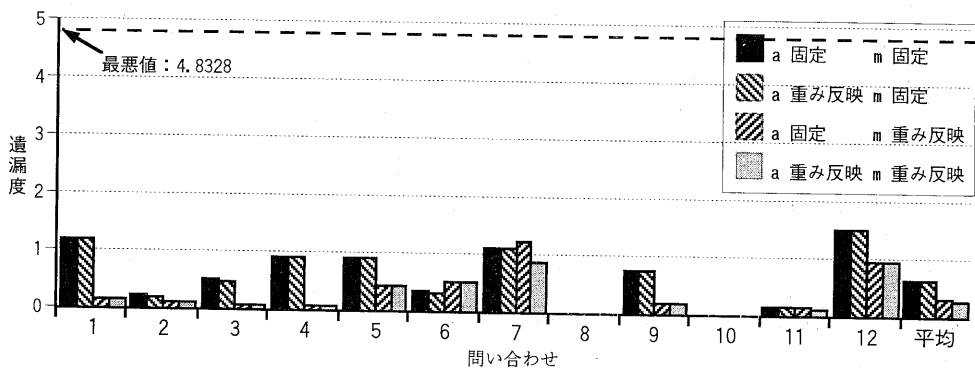


図3 各問い合わせにおける検索結果の遺漏度

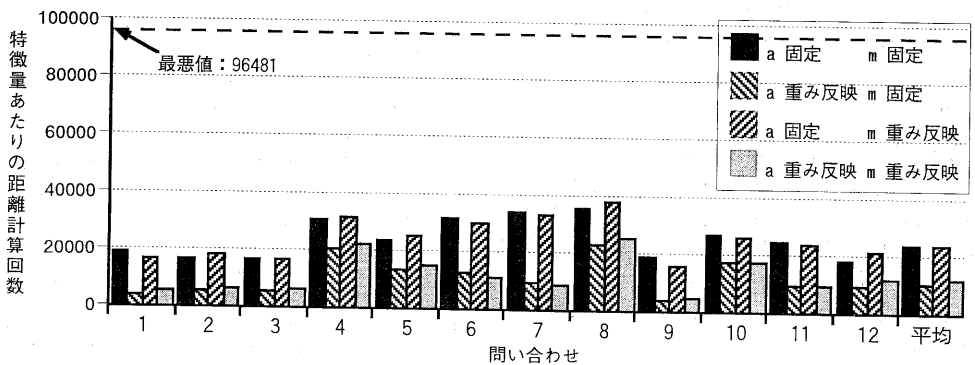


図4 各問い合わせにおける特徴量あたりの距離計算回数

図, すなわち, 例えば問い合わせ6において, 「3つの特徴量全てが同じくらい重要」なのか, それとも本当に「3つの特徴量全てが中くらいしか重要でない」のかを syntactic には識別できないため, 正規化の採用については検討中である。

図4は, 各問い合わせごとに, 距離計算回数の総和を重みが0でない(すなわち実際に検索された)特徴量の数で割った値(以下, 「特徴量あたりの距離計算回数」と呼ぶ)を示したものである。参考として最悪値, すなわちインデックスを利用せず全件に対して距離を計算した場合の計算回数(i.e. 全データ数 = 96481)も示してある。

この図からは, 以下の2つが読み取れる。

- ・ a 重み反映法は計算回数削減に有効
- ・ m 重み反映法は計算回数を多少増加させる

近似係数 (a) 重み反映法の有効性については既報でも述べたが, 新たに, 余裕係数 (m) 重み反映法と組み合わせてもこの有効性に変わることがわかった。「 m 重み反映」について「 a 固定」と「 a 重み反映」を比較すると, 最低約28%, 最高約73%, 平均約52%の計算回数削減を達成した。

一方, 余裕係数重み反映法は, 概して計算回数を多少増加させる方向に働くことが見て取れる。しかし, 平均で約9%と小幅の増加にとどまっております, この数値は事前に予想したより小さい。予想では, 問い合わせおよび特徴量によっては, 従来 $m = 6$ であった余裕係数が, 最大 $m_i = 21$ となる(例えば問い合わせ1の色相)こともあるため, 影響はもう少し大きくなると考えていた。

最後に, 近似係数重み反映法と余裕係数重み反映法の組み合わせについて見ると, 遺漏度/距離計算回数とも, 従来法に比べてかなり低減できている。したがって, 提案した2つの最適化手法の組み合わせは, 検索の正確さ/検索コスト両方を同時に改善する非常に有効な方法といえる。

5. まとめ

複数特徴量の重みつき類似検索において, 最終的な検索の正確さ/効率を向上させるため, 与えられた重みに基づいて各特徴量への検索を最適化する「近似係数重み反映法」「余裕係数重み反映法」の2つの手法について述べた。

そして, 既報[1]で評価した「近似係数重み反映法」に加えて, 新たに「余裕係数重み反映法」および両者の組み合わせ手法について新たに評価

実験を行い, 「余裕係数重み反映法」が検索の正確さの向上に有効であること, および, 組み合わせ手法により検索の正確さと検索効率が同時に改善されることを示した。

今後は,

- ・ 近似係数 a , 余裕係数 m を決める関数 f_a, f_m の更なる検討
- ・ さまざまな多次元木構造インデックスでの評価(ノードサイズの変更, 異なる木構造 C-tree [6] の利用等)
- ・ 実験規模の拡大

等を中心に研究を進める予定である。

参考文献

- [1] 谷口, 山室, "画像内オブジェクトの自動抽出を使った画像検索システム - ExSight -," 信学会, 第8回データ工学ワークショップ 1997.
- [2] 赤間, 紺谷, 三井, 串間, "画像内オブジェクトの自動抽出を使った画像検索システム ExSight," 信学, 第8回データ工学ワークショップ 1997.
- [3] Curtis, Taniguchi, Nakagawa, and Yamamuro, "A Comprehensive Image Similarity Retrieval System that utilizes Multiple Feature Vectors in High Dimensional Space," 情処 DBS-113 1997.
- [4] Konya, Kushima, "A Rotation Invariant Shape Representation Based Wavelet Transform," British Computer Society 'The Challenge of Image Retrieval' Workshop 1998.
- [5] Taniguchi, Yamamuro, "Multiple Inverted Array Structure for Similar Image Retrieval," Proc. of IEEE Intl. Conf. on Multimedia Computing Systems (ICMCS) 1998.
- [6] Curtis, Nakagawa, Taniguchi, Yamamuro, "Similarity Indexing in High Dimensional Image Space," 情処 DPS-82 1997.
- [7] White, Jain, "Similarity Indexing: Algorithms and Performance," Proc. of SPIE: Storage and Retrieval for Image and Video Databases IV 1996.
- [8] Arya, Mount, Netanyahu, Silverman, Wu, "An Optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions," Proc. of Fifth Annual ACM-SIAM Symposium on Discrete Algorithms 1994.
- [9] Flickner, Sawhney, Niblack, et al, "Query by Image and Video Content: The QBIC System," IEEE Computer Sept. 1995:23-32.
- [10] Gupta, Jain, "Visual information retrieval," Communications of the ACM vol. 40 No. 5 (May 1997):71-79
- [11] Guttman, "R-trees: a dynamic index structure for spatial searching," Proc. of the ACM SIGMOD Intl. Conf. on the Management of Data 1984.