画像の類似検索における最近接点の有意性に関する考察

片山紀生　佐藤真一

国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
TEL (03) 4212-2620　E-mail {katayama,satoh}@nii.ac.jp

キーワード：最近接点の有意性、マルチメディア情報、類似検索、最近接点探索

特徴量空間における最近接点探索は、マルチメディア情報の類似検索を実現する手段として、広く使われている。一方、データベースシステム分野の最近の研究成果から、高次元空間では、距離に有意な差が生じず、多数の点が同程度の距離に存在する状態が起こり得ることが明らかになっている。そこで、本論文では、最近接点の有意性とマルチメディア情報の類似性との関係を明らかにするために、画像検索を対象として評価実験を試みる。実験結果によると、最近接点の有意性は検索結果の類似性と密接な関係を持っており、検索結果の有意性の判定にも効果があることが明らかになった。

Effectiveness of the Nearest Neighbor's Significance in the Similarity Retrieval of Images

Norio Katayama　Shin'ichi Satoh

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo 101-8430, Japan
TEL +81-3-4212-2620　E-mail {katayama,satoh}@nii.ac.jp

Keyword : Significance of Nearest Neighbors, Multimedia Information, Similarity Retrieval, Nearest Neighbor Search

Nearest-neighbor (NN) search in high dimensional space is widely used for the similarity retrieval of images. Recent research results in the literature reveal that NN-search might return insignificant NNs in high dimensional space because points could be so scattered that every distance between them might yield no significant difference. Hence, we devised a way to estimate the significance of NNs based on the local intrinsic dimensionality. In this paper, we evaluate the proposed method with applying it to the similarity retrieval of natural photo images. The experimental result demonstrates the effectiveness of the proposed method.

1 はじめに

特徴量空間における最近接点探索は、マルチメディア情報の類似検索を実現する手段として、広く使われている。この手法では、個々のマルチメディア情報から種々の特徴量を抽出し、マルチメディア情報を多次元空間中のベクトルに対応付ける。そして、ベクトル間の距離によって類似度を判定し、類似検索を実現する。特徴量空間の重要な性質のひとつは、高次元であることである。例えば、カラーヒストグラムの場合、しばしば 20 次元以上の特徴量ベクトルが使われる。

一方、データベースシステム分野の最近の研究成果から、高次元空間では、低次元空間では想像できないような興味深い現象が起こることが明らかになっている。高次元空間の自由度があまりにも高いために点が散在してしまい、点相互の距離に有意な差が生じないことが起こり得るのである。典型的な例は、単位超立方体中に点が一様に分布している場合であり、片山ら [1, 2] は、それらの点の相互の距離にほとんど差がないことを実験的に示している。そして最近、Beyer ら [3] によって、一様分布よりも広い条件のもとでも、次元が高くなるにつれて、最近接点までの距離が最遠点までの距離に漸近することが示されている。

このように高次元空間では、距離に有意な差が生じず、多数の点が同程度の距離に存在する状態、いわば、どんぐりの背比べと呼ぶべき状態が起こる。このような状態では、最近接点探索を行っても、有意な結果は得られないことが予想される。なぜならば、同程度の距離に多数の点が存在するため、近いものとそうでないものとの間に有意な差が生じないからである。そして、最近接点探索の結果によってマルチメディア情報の類似性を判別しようとしても、有意な結果は得られないことになる。

以上のように、マルチメディア情報の類似性を特徴量空間での最近接点探索によって判別しようとする場合、最近接点の有意性に注意しなければならないことがわかる。そこで、我々は、これまでに、局所的な埋め込み次元数に基づいて最近接点の有意性を判定する方法を考案し、画像検索への適用を進めてきた [4, 5]。この手法では、最近接点と同定度の距離に存在する点を数えることによって、最近接点の有意性を判定する。これまでの実験から、この手法が、最近接点の有意性の判定に効果があることは明らかになっている。しかし、最近接点の有意性がマルチメディア情報の類似性とどれほど密接に関係しているのかについては、未だに明らかになっていない。

そこで、本論文では、最近接点の有意性とマルチメディア情報の類似性との関係を明らかにするために、画像の類似検索を対象として評価を試みる。デジタル写真画像のコレクションとしてよく知られている Corel 社の製品を実験データとして使用し、すべての画像を質問画像として最近接点探索を実行した。そして、我々が提案している手法により、最近接点の有意性を判定し、検索結果の類似性と最近接点の有意性との間に、どのような関係があるのか検証した。その結果、最近接点の有意性は検索結果の類似性と密接な関係を持っており、最近接点の有意性が高いほど、検索結果としての有意性も高いことが明らかになった。

この論文の構成は、以下のとおりである。2節では、高次元空間において有意性の低い最近接点が現れ得ることを示す。3 節では、最近接点の有意性を評価する手法について説明する。そして、評価実験の結果を 4 節で示し、5 節で結論を述べる。

2 最近接点の有意性

最近接点探索を使う場合、我々は、見つかった最近接点が、他の点よりもはるかに近くにあることを期待する。しかし、この直感は、高次元空間では必ずしも成立しない。例えば、点が単位超立方体の中に一様に分布している場合、二点間の距離は、あらゆる組合せについて見ても、ほとんど差が生じないことがある。図 1は、100,000 個の点を一様に単位超立方体の中に生成し、あらゆる組合せについて距離を求め、その最小値、平均値、最大値を示したものである。図が示すとおり、次元が高くなるにつれて、距離の最小値が著しく増加しており、最大値に対する最小値の比は、16 次元で 24%、32 次元で 40%、64 次元で 53% になっている。したがって、64 次元空間では、最近接点までの距離が、最も遠い点までの距離の 53% 以上になってしまっているのである。このような場合、我々は、これらの最近接点を、有意性の低い最近接点と見做すことが可能である。なぜならば、最近接点と他の点との差は無視できるほど小さく、他の点は、最近接点と同程度に、質問点（最近接点探索の基準となる点）の近くに存在しているからである。

図 1が示すとおり、有意性の低い最近接点は、次元が高くなるほど起こりやすくなる。この特性は、k 番めの最近接点までの距離の期待値を求めることによって確認できる。N 個の点が、質問点を中心とする超球の中に一様に分布しているとき、k 番めの最近接点までの距離の期待値 d_{kNN} は、次式のようになる [6]:

$$E\{d_{kNN}\} \approx \frac{\Gamma(k+1/n)}{\Gamma(k)} \frac{\Gamma(N+1)}{\Gamma(N+1+1/n)} r, \quad (1)$$

ここに、n は次元数であり、r は超球の半径である。そ

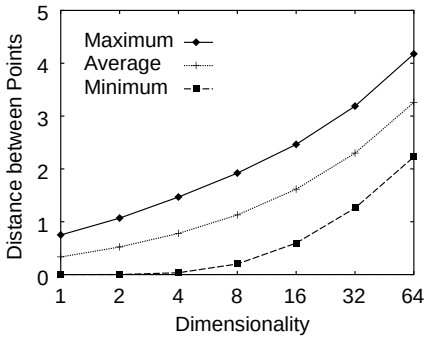


図 1: 単位超立方体に一様に生成した 100,000 個の点の間の距離。([2] から引用).

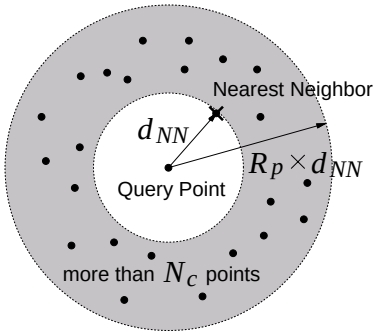


図 2: 有意性の低い最近接点の定義。

して、k 番めの最近接点までの距離と $k+1$ 番めの最近接点までの距離の比は、次式のようになる [6]:

$$\frac{E\{d_{(k+1)NN}\}}{E\{d_{kNN}\}} \approx 1 + \frac{1}{kn}. \qquad (2)$$

このように、質問点の周囲に点が一様に分布している場合、k 番めの最近接点と $k+1$ 番めの最近接点の比は、次元が高くなるほど小さくなることが期待されるのである。これは、有意性の低い最近接点が、低次元空間よりも高次元空間で起こりやすいことを示している。

3 最近接点の有意性の評価法

3.1 有意性の低い最近接点の定義

既に述べたとおり、有意性の低い最近接点は高次元空間で起こりやすく、類似検索にとって厄介な存在となる。しかし、だからと言って、高次元空間が、類似検索に役立たないということではない。実世界のアプリケーションでは、データの分布に偏りがあるため、埋め込み次元数（有効な次元数）は、特徴空間の次元数よりも小さいことが期待できる。例えば、データの分布がいくつかの優勢な次元によって支配されている場合には、埋め込み次元数は、それらの優勢な次元の数になる。さらに、埋め込み次元数は、データセット全体で一貫しているとは限らず、局所的な領域ごとに異なっている可能性がある。したがって、ある領域で最近接点の有意性が低くても、別の領域では最近接点の有意性が高い場合が起こり得るのである。そこで我々は、最近接点の有意性を、局所的な埋め込み次元数に基づいて評価する手法を考案した [4, 5]。

我々は、まず、「有意性の低い最近接点」の定義として、次の定義を考案した。

定義 1 d_{NN} を質問点から最近接点までの距離とする。また、質問点からの距離が d_{NN} から $R_p \times d_{NN}$ の範囲にある領域を、質問点の近傍領域と呼ぶことにする。このとき、近傍領域に N_c 個以上の点が存在すれば、その最近接点を「有意性の低い最近接点」と見做す。

ここに、$R_p(>1)$ と $N_c(>1)$ は、制御パラメータである（図 2）。R_p は質問点の近傍領域を決定し、N_c は近傍領域の輻輳を決定する。

3.2 パラメータの設定

上記のように、R_p と N_c は、有意性の低い最近接点の定義において、本質的な役割を果たしている。ここで、埋め込み次元数に基づいて、これらのパラメータを設定する方法を示す。

まず最初に、最近接点の有意性と埋め込み次元数とを、式 (2) によって関連付ける。この式は、埋め込み次元数が n のときにも成り立つ [6]。すなわち、この式の n は、データ空間の次元数を意味しているのではなく、有効な次元数（すなわち、埋め込み次元数）を意味しているのである。式 (2) は、$d_{(k+1)NN}$ と d_{kNN} の期待値の比が、k が増加するにつれて単調減少することを示している。したがって、期待値の比が最大となるのは、次式のように 1 番めと 2 番めの最近接点においてである。

$$\max_k \frac{E\{d_{(k+1)NN}\}}{E\{d_{kNN}\}} = \frac{E\{d_{2NN}\}}{E\{d_{1NN}\}} \approx 1 + \frac{1}{n}. \qquad (3)$$

この式は、k 番めの最近接点と $k+1$ 番めの最近接点の相対差の期待値が、次元数が高くなるにつれて単調減少することを示している（図 3）。5 次元の場合には、20% 以下の差しか期待できず、10 次元の場合には、10% 以下の差しか期待できないことがわかる。このように、埋め込み次元数は、最近接点の相対的な有意性を見積もることを可能にする。

次に、埋め込み次元数 n とパラメータ R_p、N_c を関連付ける。局所的な領域において、埋め込み次元数 n で、

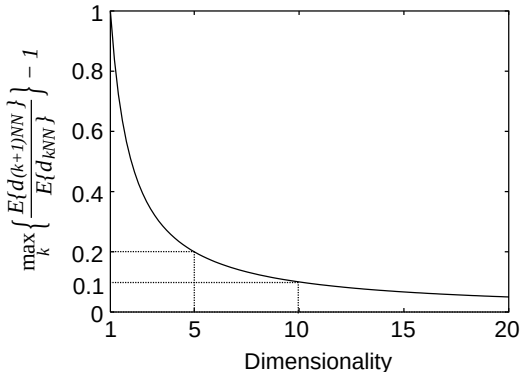


図 3: k 番めと $k+1$ 番めの最近接点の相対差の期待値。

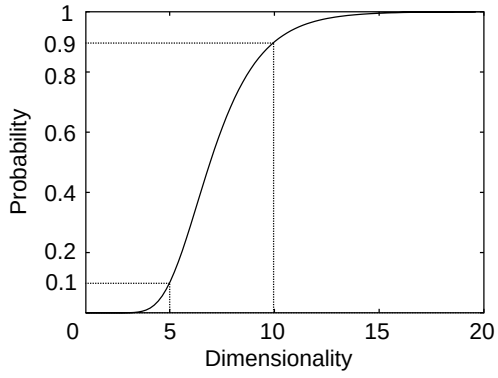


図 4: R_p が 1.84471 N_c が 48.0277 の場合の棄却率。

点が一様に分布していると仮定すると、R_p で決定される近傍領域に N_c 個以上の点が存在する確率は、次式のようになる。

$$Pr\{N_c \text{ or more in } R_p\} = (1 - (1/R_p)^n)^{N_c} \qquad (4)$$

定義 1 によれば、R_p で決定される近傍領域に N_c 個以上の点が存在すれば、最近接点の有意性が低いと見做されるので、式 (4) は、埋め込み次元数が n のときに、最近接点の有意性が低いと見做される確率を示していることになる。そこで我々は、この確率を「(最近接点の) 棄却率」と呼んでいる。棄却率は、埋め込み次元数が大きくなるにつれて単調に増加する。また、以下に示すように、二つの制御点によって、容易に制御することが可能である。今、棄却率を、次元数 ν_1 でρ_1、次元数 ν_2 で ρ_2 に設定したいとする ($0 < \rho 1 < \rho 2 < 1$ かつ $1 < \nu_1 < \nu_2$)。すると、パラメータ R_p と N_c は、次の連立方程式を解くことによって決定できる。

$$(1 - (1/R_p)^{\nu_1})^{N_c} = \rho_1 \qquad (5)$$

$$(1 - (1/R_p)^{\nu_2})^{N_c} = \rho_2 \qquad (6)$$

N_c を消去することによって次式を得る。

$$\frac{\log(1 - (1/R_p)^{\nu_1})}{\log(1 - (1/R_p)^{\nu_2})} = \frac{\log \rho_1}{\log \rho_2} \qquad (7)$$

この方程式は、算術的に解くことはできない。しかし、左辺が R_p に対して単調に増加するため、ニュートン法などの数値解法によって容易に解くことができる。R_p が求まれば、次式によって N_c も求まる。

$$N_c = \frac{\log \rho_1}{\log(1 - (1/R_p)^{\nu_1})}. \qquad (8)$$

以上のように、式 (7) と (8) を用いることで、二つの制御点 (ν_1, ρ_1) と (ν_2, ρ_2) から、R_p と N_c を決定することが可能になる。

次に問題になるのが、どのようにして制御点を決定するかであるが、我々は、これらの制御点を、低域通過型フィルタとのアナロジーから、カットオフ点 (ν_c, ρ_c) と遮断点 (ν_r, ρ_r) のペアとして設定する方法を提案している。$n < \nu_c$ の範囲は、通過域に相当し、埋め込み次元数がこの範囲にあると、高い確率で有意性が高いと見做される。一方、$n > \nu_r$ の範囲は、遮断域に相当し、埋め込み次元数がこの範囲にあると、高い確率で有意性が低いと見做される。$\nu_c < n < \nu_r$ の範囲は、過渡域に相当する。また、我々は、ν_c と ν_r を k 番めと $k+1$ 番めの最近接点の相対差の近似的な期待値（式 (3) と図 3）によって決定することを提案している。例えば、5 次元では 20% 以下の差しか期待できず、10 次元では 10% 以下の差しか期待できないので、カットオフ次元数を 5、遮断次元数を 10 とすることが考えられる。そして、カットオフ次元数での棄却率を 0.1、遮断次元数での棄却率を 0.9 とすると、二つの制御点は、(5, 0.1) と (10, 0.9) となる。これらの制御点から、式 (7) と (8) によって、R_p と N_c を求めると、R_p が 1.84471、N_c が 48.0277 となる。図 4 は、これらのパラメータを用いたときの棄却率を、式 (4) によって計算した結果である。

4 画像の類似検索による評価

4.1 評価実験の内容

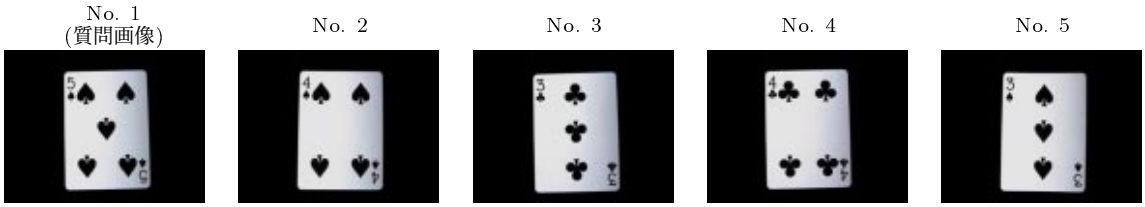
以上に述べたように、R_p で決まる近傍領域に N_c 個以上の点が存在するかどうか調べることによって、最近接点の有意性（厳密に言えば、その近傍領域が、距離的に有意な差を期待できる埋め込み次元数を持っているかどうか）を判定することができる。この節では、この手法を画像の類似検索に適用し、評価を試みる。

実験データとしては、デジタル写真画像のコレクシ

(a) テクスチャ (検索結果における有意性の高い最近接点の数: 92)

(b) 空 / 海 (検索結果における有意性の高い最近接点の数: 62)

(c) 肖像画 (検索結果における有意性の高い最近接点の数: 35)



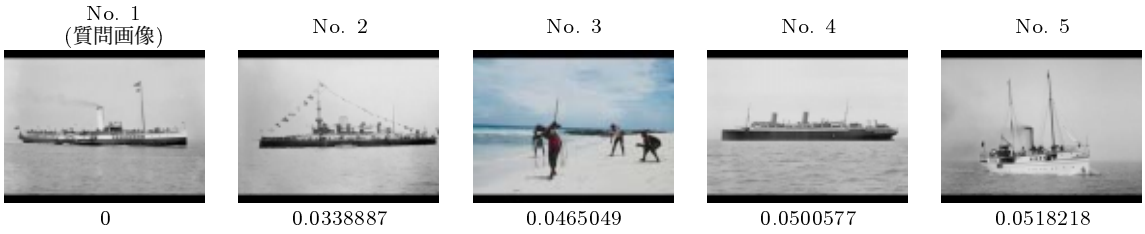
(d) トランプ (検索結果における有意性の高い最近接点の数: 23)

(e) 花火 (検索結果における有意性の高い最近接点の数: 23)

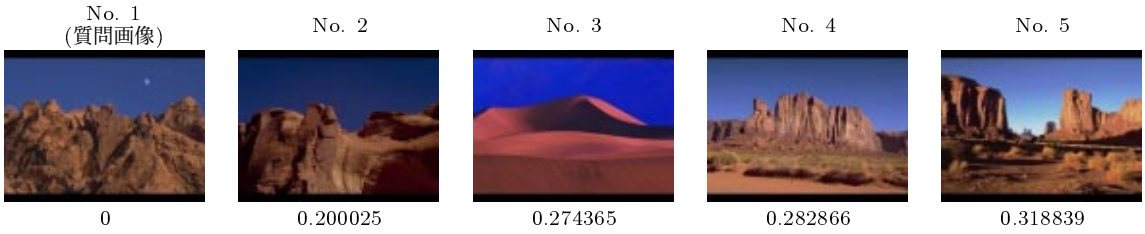
(f) 日没 (検索結果における有意性の高い最近接点の数: 19)

図 5: 検索結果の例。

(g) カンフー (検索結果における有意性の高い最近接点の数: 18)

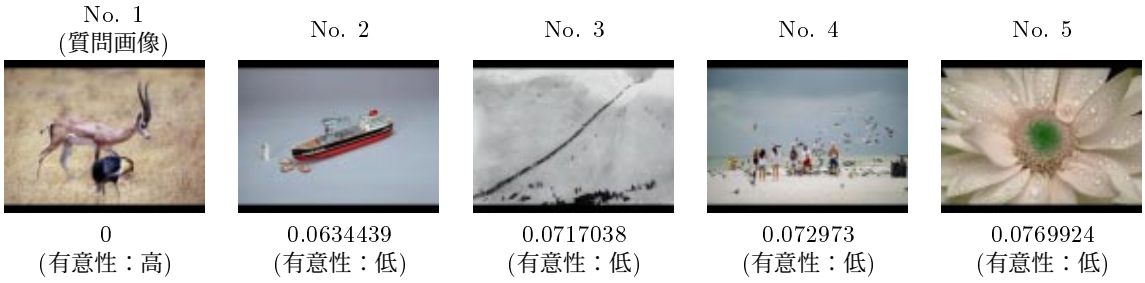


(h) 蒸気船 (検索結果における有意性の高い最近接点の数: 17)

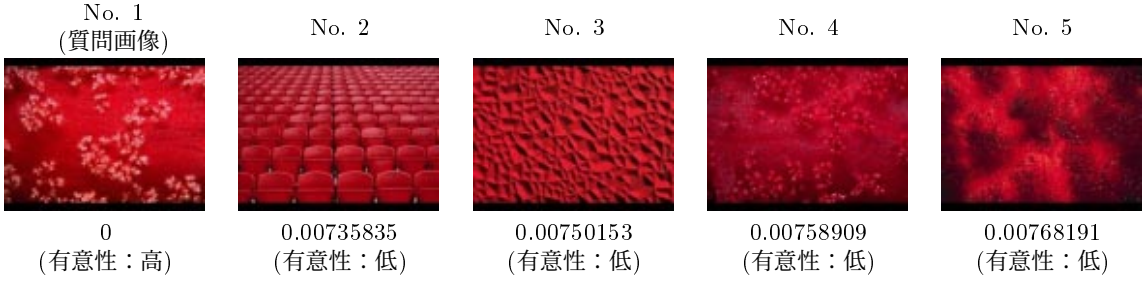


(i) 砂漠 (検索結果における有意性の高い最近接点の数: 16)

図 5: 検索結果の例 (つづき)。



(a) 類似した画像が得られない例



(b) 差の小さい画像しか得られない例

図 6: 質問点以外に有意性の高い最近接点が得られなかった場合の例

表 1: 検索結果における有意性の高い最近接点の数。

検索結果における有意性の高い最近接点の数	度数
100	0
90 ～ 99	6
80 ～ 89	40
70 ～ 79	0
60 ～ 69	1
50 ～ 59	53
40 ～ 49	13
30 ～ 39	37
20 ～ 29	87
10 ～ 19	269
2 ～ 9	13649
1	46040
合計	60195

表 2: 検索結果の種類。

画像の種類 (人手による分類)	検索結果における有意性の高い最近接点の数 (最大値)
テクスチャ	92
空 / 海	62
肖像画	35
トランプ	23
花火	23
日没	19
カンフー	18
蒸気船	17
砂漠	16

ョンとしてよく知られている Corel 社の Corel Photo Collection を使用した。Corel Gallery 1,000,000 という製品に含まれている 60,195 枚の画像を実験データとした。画像の特徴量としては、カラーヒストグラムを使用した。色空間としては、マンセル表色系を使い、9 色の部分空間（黒、灰、白、ならびに、色相で分割した 6 色）に分割した。画像の構図を反映させるために、個々の画像を 4 分割し、それぞれの領域についてカラーヒストグラムを求め、それらを結合して 36 次元の特徴ベクトルを算出した。特徴ベクトル間の類似度は、ユークリッド距離によって評価した。全ての画像の特徴ベクトルを質問点として、100 番めまでの最近接点を探索する処理を実行した。そして、それら 100 個の最近接点について、上記の方法で有意性を評価した。実験データ中の画像を質問点としているので、探索によって見つかった最近接点のひとつは、質問点そのものである。

4.2 検索結果における有意性の高い最近接点の数

表 1 は、検索結果中に含まれる有意性の高い最近接点の数を示したものである。実験データ中の全ての画像を質問画像としているので、度数の合計は、実験データ中の画像と同じになっている。また、我々は、有意性の高い最近接点が比較的多く含まれている検索結果について、どのような種類の画像が得られているのか人手によって分類した。その結果を、表 2 に示す。図 5は、分類された検索結果の例を示したものである。紙面の制約のため、上位 5 件までを示している。個々の画像の下の数値は、質問点からの距離を示している。いずれの例についても、類似性の高い結果が得られていることがわかる。

表 1 からわかることとして興味深いのは、60,195 枚のうちの 46,040 枚（76.5%）の画像からは、有意性の高い最近接点が、わずか一つしか得られていないことである。検索結果中には、質問画像も含まれているため、この見つかったものは、質問画像そのものに過ぎない。従って、質問画像以外には、有意性の高い最近接点がひとつも見つかっていないことになる。この場合、質問画像は、ほぼ同定度の類似性を持つ多数の画像に囲まれていることになる。このような現象が、46,040 枚もの画像について起こるのは、奇妙なことのようにも思えるが、Corel のコレクションは、多様な写真画像を収集しているため、似通った画像が存在しなくても不自然ではないと考えられる。さらに、このコレクションには、テクスチャ画像も含まれており、この場合、わずかな差しかない画像が多数存在し、有意性の高い結果が得られないことになる。図 6 は、質問画像以外に有意性の高い最近接点が見つからなかった場合の例である。紙面の制約のため、上位 5 件までを示している。図 6 (a) は、質問画像が多くの類似していない画像によって囲まれている場合であり、図 6 (b) は、わずかな差しかない多数の画像によって囲まれている場合である。

以上の結果から、最近接点の有意性は、検索結果の類似性と密接な関係を持っていることがわかる。そして、最近接点の有意性が高いほど、検索結果としての有意性も高く、最近接点の有意性を調べることによって、検索結果がどれほどの有意さで類似しているのかを評価できることがわかる。このような類似性の評価法は、特に、人間への検索結果の呈示を伴う対話的な検索システムに

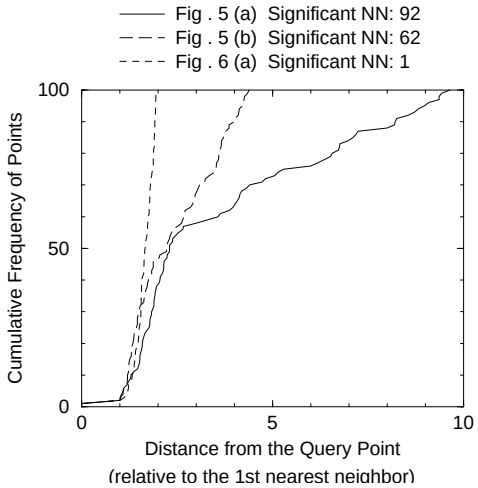


図 7: 質問点の周囲に存在する点の累積分布。

おいて有効であると考えられる。

4.3 質問点の周囲に存在する点の累積分布

最近接点の有意性の判定結果の妥当性を検証するために、我々は、質問点の周囲に存在する点の累積分布を求めた。図 7 は、図 5 (a)、図 5 (b)、図 6 (a) の検索結果について、質問点からの距離に関する累積分布を示したものである。横軸は、質問点からの距離を表しており、それぞれの検索結果の 1 番めの最近接点までの距離で正規化された値になっている。一方、縦軸は、質問点の周囲に存在する点の累積分布（度数を累積したもの）を示している。図からわかるとおり、質問点の周囲の点の分布は、質問点によって大きく変化している。有意性の高い最近接点が多く見つかった検索結果（図 5 (a) と (b)）は、ひとつしか見つからなかった検索結果（図 6 (a)）に比べて、点の分布が広くなっており、多数の点が同定度の距離に存在する状態を検出したいという提案手法の狙いに合致した結果となっている。

5 むすび

本論文では、高次元空間における最近接点の有意性に着目し、有意性の高低とマルチメディア情報の類似性との間にどのような関係があるのか評価した。画像の類似検索を例として実験を行った結果、最近接点の有意性は検索結果の類似性と密接な関係を持っており、最近接点の有意性が高いほど、検索結果としての有意性も高いことが明らかになった。最近接点の有意性とマルチメディア情報の類似性に密接な関係がある場合、検索結果がどれほど類似しているのかを最近接点の有意性に基づいて評価することが可能になる。このような類似性の評価法は、特に、人間への検索結果の呈示を伴う対話的な検索システムにおいて有効であると考えられる。本論文の実験結果は、その可能性を支持する結果となっている。

謝辞

本研究の一部は、文部省創成的基礎研究費（09NP1401）、および、文部省科学研究費補助金（奨励 12780251）の補助を受けた。

参考文献

[1] N. Katayama and S. Satoh, “The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries,” Proc. of the 1997 ACM SIGMOD, Tucson, USA (May 1997) pp. 369–380.

[2] 片山紀生，佐藤真一，“SR-Tree: 高次元点データに対する最近接検索のためのインデックス構造の提案,” 信学論（D-I），vol.J80-D-I，no.8，pp.703–717，Aug. 1997.

[3] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When Is “Nearest Neighbor” Meaningful?,” Proc. of the 7th Int. Conf. on Database Theory, Jerusalem, Israel, pp.217–235, Jan. 1999.

[4] N. Katayama and S. Satoh, “Significance-Sensitive Nearest-Neighbour Search for Efficient Similarity Retrieval of Multimedia Information,” First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM’99), Orlando, USA (Oct. 1999).
(available on-line at http://www.info.uqam.ca/~misrm/papers/katayama.ps.gz)

[5] 片山紀生, 佐藤真一, “最近接点の有意性の評価によるマルチメディア情報の効率的な検索法,” 電子情報通信学会技術研究報告, DE2000-83, 2000.

[6] K. Fukunaga, “Introduction to Statistical Pattern Recognition (2nd ed.),” Academic Press, 1990.