様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　２５　日現在

機関番号：　２０１０３
研究種目：　若手研究(A)
研究期間：　2011～2014
課題番号：　２３６８０００８

研究課題名（和文）高次元特徴量ベクトルの最近傍探索を行う改良型ＬＳＨアルゴリズムの研究

研究課題名（英文）Improved LSH Algorithm for Approximate Nearest Neighbor Search of High Dimensional Vectors

研究代表者

寺沢　憲吾（Terasawa, Kengo）

公立はこだて未来大学・システム情報科学部・准教授

研究者番号：１０４３５９８５

交付決定額（研究期間全体）：（直接経費）　　15,100,000円

研究成果の概要（和文）：本研究は、大量に蓄積された高次元ベクトルデータから、問い合わせデータに最も近いデータを高速に探索する最近傍探索アルゴリズムを開発することを目的としている。1000万画像からなるデータベースに対し、ユークリッド距離を用いる場合にはSLSH(Spherical LSH)、インタセクション類似度を用いる場合には改良版LSHをそれぞれ適用し、一定の成果を得た。

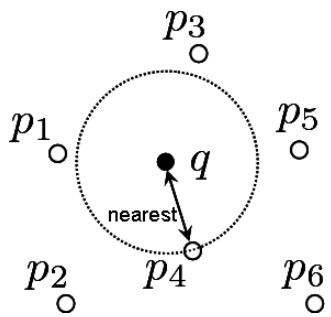
研究成果の概要（英文）：In this study, we aim to establish an algorithm to find the nearest-neighbor of the query vector among the huge database of high-dimensional vectors. We applied the SLSH (Spherical LSH) based method for the case of Euclid distance and the improved LSH based method for the intersection similarity, and confirmed that our method is efficient than existing methods.

研究分野：　画像処理、パターン認識、情報検索、アルゴリズム

キーワード：　アルゴリズム　画像、文章、音声等認識　コンテンツ・アーカイブ

１．研究開始当初の背景

情報通信技術の発達に伴い、画像、動画、音声などの多様なデータが市井に氾濫するようになり、大量のデータから必要とするものを適切に選択する技術の重要性が増している。データ集合の中から問い合わせデータに最も近いデータを探索する最近傍探索問題（図１）は、そのための基本となる問題であるが、対象となる全データに対して条件に対する適合性を評価する全探査の手法は、対象となるデータのサイズに比例する計算コストがかかるため、大量データを扱う場合は現実的に有用ではない。大量データに対して効率的な最近傍探索アルゴリズムとして kd-tree や R-tree のような空間インデックスを用いたデータ構造を用いる方法があるが、これらは「次元の呪い」を受けることが知られている。すなわち、ベクトル空間の次元 d が大きくなるにつれ、d に対し指数関数的な処理（前処理時間やメモリサイズ）を要求するか、あるいは計算コストが全探査なみに収束してしまうかのいずれかに陥る。パターン認識に用いる特徴量ベクトルは数百から数千、数万次元のものがあり、「次元の呪い」の影響を受ける kd-tree や R-tree のような方法は現実的に有用ではない。



（図１）最近傍探索問題

そこで近年では近似アルゴリズムの有用性が注目されている。中でも Locality - Sensitive Hashing (LSH, 参考文献[1]) は注目されており、学術雑誌 Communications of the ACM (CACM) の 2008 年 1 月号における特集記事“ Breakthrough Research ”（参考文献[2]）の中で、画期的な研究事例 2 件のうちの 1 つとして取り上げられている。また、パターン認識分野において、LSH アルゴリズムを用いた応用研究が盛んに行われつつある。

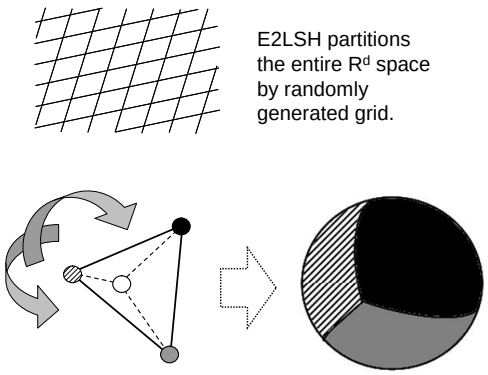
研究代表者の寺沢は 2007 年に、この LSH の部分的改良手法として、Spherical LSH（SLSH, 参考文献[3]）を提案している。これは、用いるデータ集合と距離尺度のいずれについても極力一般性を持たせようとした LSH に対し、パターン認識において解くべき問題の実際に鑑み、距離をユークリッド距離に限定し、しかも対象ベクトルを単位ベクトルに限るという仮定を持たせることで、LSH よりも効率的な近傍探索が可能であることを示したものである。この SLSH は前述の CACM の特集でも LSH のバリエーションの一つとして引用され、また、性能についてさらに詳細な検証を施した論文（参考文献[4]）は平成 21 年度電子情報通信学会論文賞を受賞するなど、高い評価を得た。

２．研究の目的

前述の通り、最近傍探索問題の近似解法は近年最も注目されている研究課題の一つである。本研究では、既に開発している SLSH に改良を加え、計算量と必要メモリ量の両面で計算コストの縮減を目指すとともに、これまで以上に広範なパターン認識の問題に応用できるよう、ユークリッド距離以外の類似尺度に対して適用可能に拡張すべく、新たなハッシュ関数の開発を目指す。一例として、前処理の際に次元数 d の 3 乗に比例する計算コストが必要であるため、実用上次元数 d は数百程度が限界であった点などの解消を目指す。本研究により高次元特徴量ベクトルの最近傍探索のコストが低下すれば、その益は画像認識や文字認識にとどまらず、言語処理やデータマイニングなどの分野にも及び、各種応用手法の効率化や新たな研究の促進という結果につながることが期待できる。

３．研究の方法

従来の LSH では、d 次元空間を超平面により格子状に分割することでハッシュ関数を作成し、同じハッシュ値を持つ点を探索対象にするという方法がとられていた（図 2 上段）。一方、SLSH では、d 次元単位球上にランダムに回転させた正多胞体を配置し、その頂点をシーズ点として、ボロノイ分割により単位球表面を分割することで、従来の LSH よりも効率の高い最近傍検索を行うことができる（図 2 下段）。



（図 2）LSH における空間分割法と SLSH における超球表面分割法。図は 3 次元空間の例だが、実際は d 次元空間を考える。

本研究ではまず、実際の画像データセットから得られた特徴量を用いて、検索精度を確認するとともに、前処理部分と検索本体部分のそれぞれについて、設定パラメータによる計算コスト（実測時間）の変化度合いが理論推計通りの挙動を示すかを確認する。次いで、大容量メモリ搭載のコンピュータを用いて実測時間を測定し、理論推計値と同様の結果が得られるかどうか検証する。

これらの結果をふまえ、従来用いていた、正多胞体の頂点に基づくハッシュ関数とは別の、新たなハッシュ関数の設計開発を進める。

４．研究成果

(1) 実際の画像データセットから得られた特徴量を用いて、検索精度を確認するとともに、前処理部分と検索本体部分のそれぞれについて、設定パラメータによる計算コスト（実測時間）の変化度合いが理論推計通りの挙動を示すかを検証する。

既往の研究（参考文献[4]）の段階では検証実験には主に一様分布のランダムデータを用いていた。一方、実際のパターン認識の特徴量は特徴空間内に一様に分布しているわけではなく、偏りをもって分布していると考えられる。本検証では、実画像データベースから特徴量を作成し、こうした特徴量に対しても SLSH が有効に機能することを確認する。

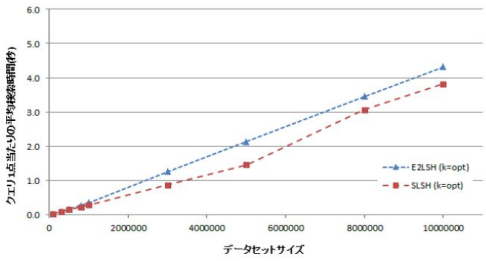
データ集合としては、"80 million tiny images" （参考文献[5]）で公開されている 32×32 サイズの画像を用い、この画像全体をパッチとして SIFT-descriptor （参考文献[6],[7]）を用いて 128 次元の特徴量を構成した。また、問合せ（クエリ）画像としては、データベース内に存在する画像を微小に変形させたものとして、元画像の上下各 2 ピクセルをトリムした 28×28 サイズの画像を用意した。

これらの手法とデータを用い、まず 100 万画像に対して実験を行い、理論推計通りの挙動を示すことを確認し、国際会議 IWAIT2012 において報告した。

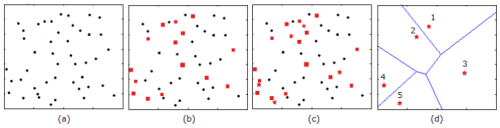
(2) さらに大規模なデータベースに対し、実測時間を測定し、理論推計値と同様の結果が得られるかどうか検証する。

前項の検証では検証環境の制約により、検証データ数の上限は 100 万程度としていたが、その後環境構築を進め、データ数を 1000 万まで増やした検証を可能にした。

図 3 はその結果である。500 万画像程度までは理論通りの挙動を示すが、データサイズがそれを上回ると、理論上最適なパラメータと、メモリサイズの制約内で最適なパラメータに乖離が発生し、SLSH の E2LSH に対する優位性が逓減することが確認された。ただしこの原因はメモリサイズの制約によるもの



（図 3）1000 万画像データベースに対する、E2LSH(青)と SLSH(赤)の計算コストの比較



（図 4）データ分布から c-means 法を用いて典型ベクトルを構成する方法

と特定されているので、将来的に計算機のメモリサイズが大きくなるか、あるいは手法の改良により必要メモリサイズが小さくなるかのいずれかにより解決されるものと予見できる。これらの結果は情報処理学会第 75 回全国大会において報告した。

(3) これらの結果をふまえ、従来用いていた、正多胞体の頂点に基づくハッシュ関数とは別の、新たなハッシュ関数の設計開発を進める。本研究ではまず実用性を重視し、一般的に良く用いられている BoF（Bag of Features, 参考文献[8]）への適用を狙い、BoF で使われることの多いインタセクション類似度に適用可能なハッシュ関数の開発を試みた。

インタセクション類似度はユークリッド距離とは異なる特性を持つため、SLSH をそのまま適用することはできない。そこで図 4 に示すように、データ分布から c-means 法を用いて典型ベクトルを構成する方法を考案し、その効果を実験により検証した。検証においては、このハッシュ関数が locality-sensitive 条件を満足することの確認と、実測時間の両面から確認を行った。その結果、この方法である程度の精度で画像検索を行うことができることを確認した。

ただし、検索精度を一定以上に保つために必要な計算コストは想定以上に大きいことが判明し、計算コストを縮減するためにはハッシュ関数にさらに工夫を加える必要があることも同時に判明した。これらの結果は deim2013 において報告した。

(4) 従来用いていた、正多胞体の頂点に基づくハッシュ関数とは別の、新たなハッシュ関数の設計開発を進めることについては、既に構築している 1000 万画像からなるデータベースに対し、過去に適用していた SLSH を上回る効率を持つハッシュ関数の開発を目的とし、いくつかのパターンで実験を試みるとともに、期待したような実験結果が得られない原因を調査すべく実装の詳細や計算環境の再確認を行ったが、研究期間内には期待したような成果を得るに至らなかった。ただし、(1),(2)で述べたように、実際にそのようなハッシュ関数を開発することができる可能性は予見しているので、本研究期間終了後も引き続き本研究課題については研究を続けていく予定である。

＜参考文献＞

[1] A. Andoni, P. Indyk, "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions," in Proc. FOCS'06, 2006.

[2] "Breakthrough research: a preview of things to come", Communications of the ACM, vol.51, issue 1, pp.104-122, 2008.

[3] K. Terasawa, Y. Tanaka, "Spherical LSH for Approximate Nearest Neighbor Search on Unit Hypersphere," in Proc. WADS2007, 2007.

[4] K. Terasawa, Y. Tanaka, "Approximate Nearest Neighbor Search for a Dataset of Normalized Vectors", IEICE Transactions on Information and Systems, vol.E92-D, No.9, 2009.

[5] A. Torralba, R. Fergus and W. T. Freeman, 80 million tiny images: a large dataset for non-parametric object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30(11), pp. 1958-1970, Nov. 2008.

[6] D. G. Lowe, Object recognition from local scale-invariant features, Proc. 7th International Conference on Computer Vision, ICCV'99, vol. 2, pp. 1150–1157, 1999.

[7] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[8] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual Categorization with Bags of Keypoints," In ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

５．主な発表論文等

〔学会発表〕（計　３　件）

① 重村拓也，清水大輝，寺沢憲吾，LSH による大規模画像データからの高速類似検索，情報処理学会第 75 回全国大会，2013 年 3 月 6 日，東北大学川内キャンパス（宮城県仙台市）．

② 清水大輝，寺沢憲吾，BoF の類似度に適合する改良版 LSH を用いた高速類似画像検索，第 5 回データ工学と情報マネジメントに関するフォーラム（deim2013），2013 年 3 月 5 日，ホテル華の湯（福島県郡山市）．

③ Hiroki Shimizu and Kengo Terasawa, Effectiveness of image searching with LSH algorithms, International Workshop on Advanced Image Technology, IWAIT2012, pp.712-717, Jan. 9-10, 2012, Ho Chi Minh City, Vietnam.

６．研究組織

(1)研究代表者

寺沢　憲吾　（TERASAWA, Kengo）

公立はこだて未来大学・システム情報科学部・准教授

研究者番号：１０４３５９８５