I-040

縮退特徴量を用いた疑似クラスタリングによる高次元近接点探索の高速化
Acceleration of High-Dimensional Nearest Neighbor Search by Quasi Clustering Using Reduced Features

山岸 史典† 片山 紀生‡ 佐藤 真一‡ 坂内 正夫‡
Fuminori YAMAGISHI Norio KATAYAMA Shin'ichi SATOH Masao SAKAUCHI

1. はじめに

数万次元以上の高次元空間における近接点探索は、ゲノム解析やテキスト処理などはじめとする様々な分野で利用されており、効率的な高次元近接点探索はいずれの分野でも切望されている。

一般に近接点探索の高速化のためにはデータの分布の偏りに基づき近接点同士のまとまりを構成しておく、クラスタリング手法がよく用いられる。クラスタリングにより、複数の点をまとめて近接性の判定ができることで距離比較の回数を削減できると期待できるためである。例えば木やハッシュは実質的に探索対象空間をクラスタリングし、対象点の高速探索を可能にしている構造であると説明できる。しかし、数万次元以上の高次元データのクラスタリングにはデータの次元の高さのために１回の距離比較計算のコストが大きく、通常多数回の距離比較が必要となるクラスタリング処理そのものが困難になる。

一方これとは別に、高次元空間内での近接点探索の高速化の一般的な手法として、より低次元の特徴量に変換し、その特徴量により探索対象点を絞り込み、高次元距離比較回数を削減する手法 [1] が知られているが、絞り込まれた後は高次元での距離比較を各点について順次行うため効果が限定的であった。

本稿では、高次元空間でのクラスタリングを併用し、より低次元の特徴量空間でクラスタリングを行い、その結果を流用して元の高次元空間をさらにクラスタリングし、２つのクラスタリング情報を同時に利用して高次元近接点探索を高速化する手法を提案する。われわれは高次元近接点探索の一つのアプリケーションとして、映像データベース中に繰り返し現れる映像断片を探索する、同一映像断片探索を行っているが、本高速化手法を同一映像断片探索に適用する実験を行い、高速化に大きな効果があることを確認した。

2. 同一映像断片探索

同一映像断片探索とは、ある映像集合の中に複数回現れる映像の断片を探索する操作である。例えば放送映像を対象にした場合、繰り返し映像が現れる頻度はさほど高くないが、ある一定量は確かに存在し、映像データの構造化に有用であるとわれわれは考えている。

実際のニュース映像中での同一映像断片の例を図１に示す。例えばニュース映像では、同一または関連のあるニューストピック同士は資料映像を共有するため、映像情報のみを用いて、意味的な関連性の抽出が可能である [4]。さらには番組のタイトル映像や CM 映像などの検出が可能であり、放送の事実の検証や出現分布等の統計量の取得に用いることができる。

†東京大学大学院情報理工学系研究科
‡国立情報学研究所, NII

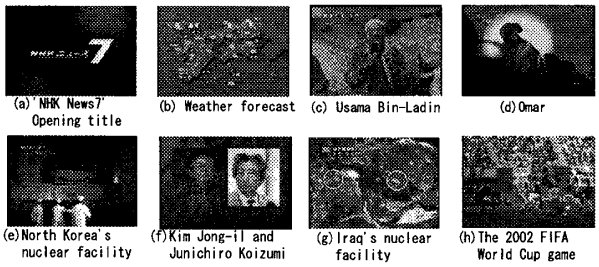

図 1: 検出された同一映像断片の例

上記例示のように、放送映像を対象として同一映像断片探索を行うことで、映像データベースの構造化や有用な情報の抽出が可能となるが、前述したように放送映像においては映像が繰り返し現れる頻度はあまり高くないので、構造化に利用可能なほどの探索結果を得るには、ある程度大量の映像を対象とすること利用価値の高い結果を得るために必要である。

われわれは映像断片探索を具体的には以下のように定義した。映像断片とは、フレーム画像の連続である映像の、時間的に連続した部分であるとする。取得された映像をその映像を構成する画像の集合としてとらえる。それぞれのフレーム画像はその集合の要素となる。その集合中で同一の映像同士を探索する。ただし、同一の画像の定義として「画像の画素値同士が互いにすべて一致する」を採用すると、放送映像取得の際のノイズ等により元々同一の画像として放送されたものでも同一と判定されないし、同一画像を用いながらも加えられたテロップの違い等は許容して同一と判定するようにしたいので、判定基準として画像間の正規化相互相関 (Normalized Cross Correlation;NCC) を用いることにした．NCC は

$$NCC(a,b) = \frac{\frac{1}{n}\sum_i^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\frac{1}{n}\sum_i^n (a_i - \bar{a})^2}\sqrt{\frac{1}{n}\sum_i^n (b_i - \bar{b})^2}} \quad (1)$$

で２枚の画像間に定義される値である。(a,b は画像,a_i,b_i は画像を構成する n 個の画素の輝度,$\bar{a}$,$\bar{b}$ は各画像の画素の輝度の平均値である．) われわれは正規化相互相関の値が 0.90 以上のものを同一の画像と定義している。この基準で通常の放送映像で同一映像断片がほぼもれなく抽出できていることを確認している。

この探索によると、対象映像アーカイブ中での同一の画像同士の出現位置が明らかになり、わずか１フレームのものも含め任意の長さの映像断片の出現を検出可能である。

映像断片探索は、高次元空間上での近接点探索と等価である。以下にその理由を説明する。

映像を構成する各フレーム画像をフレームごとに正規化し集合の要素とする。ここでの正規化とは、その各画

素値が

$$\tilde{a}_i = \frac{a_i - \bar{a}}{\sqrt{\frac{1}{n}\sum_i^n (a_i - \bar{a})^2}} \quad (2)$$

となるものとする。ここで $\tilde{a}$ は正規化画像である。これによると式 (1) は、

$$NCC(a,b) = \frac{1}{n}\sum_i^n \tilde{a}_i \tilde{b}_i \quad (3)$$

となる。

一方、正規化画像間の距離 (Image Distance) を、式 (4) のように画素毎のユークリッド距離の二乗和の平均と定義する。

$$ImageDistance(\tilde{a},\tilde{b}) := \frac{1}{n}\sum_i^n (\tilde{a}_i - \tilde{b}_i)^2 \quad (4)$$

この距離を本稿では画像間距離と呼ぶこととする。すると、

$$\begin{aligned} ImageDistance(\tilde{a},\tilde{b}) &:= \frac{1}{n}\sum_i^n (\tilde{a}_i - \tilde{b}_i)^2 & (5)\\ &= 2 - \frac{2}{n}\sum_i^n \tilde{a}_i \tilde{b}_i & (6)\\ &= 2 - 2NCC(a,b) & (7) \end{aligned}$$

が成り立つ。つまり、NCC は画像間距離を用いて表現でき、すなわち、NCC が閾値 θ_{ncc} 以上 であることをもってフレーム間同一性の基準を与えると言うことは、画像間距離が $2 - 2\theta_{ncc}$ 以下であることを基準とすることと同値である。

3. 縮退特徴量を用いた疑似高次元クラスタリング

3.1 従来手法

本稿では近接点探索問題を

問題:クエリー点 q から一定の距離 θ 以内にあるデータ点をすべて探索する。

と定義することとする。これによると同一映像断片探索は $\theta = 2 - 2\theta_{ncc}$ での探索に他ならない。

元の高次元空間から計算したより低次元の縮退特徴量を用いる手法 [1] に基づいて、同一映像断片探索のために [3, ?] で提案した高次元近接点探索手法は次の通りである。

[従来手法]

まず、図3左半分に示すように、探索の対象となる高次元ベクトル空間 H が存在するものとする。H 上で定義されている距離関数を D とする。図中の小さな黒ドットはそれぞれ高次元ベクトルデータ点である。また、クエリー点は図の左半分の中央付近に示した。従って問題の探索対象点は図中で灰色で示した円内にある黒ドットである。

まず探索に先立つ準備として、ある適当な特徴量関数 F を用い、H 上のすべての点について特徴量を計算しておき、特徴量空間 L を得ておく。F は H 上の高次元データから、より低次元の特徴量ベクトルを計算する関数である。この操作により、図中では右側に示した特徴量空間 L に H 上の各点が移される。一部の点について対応関係を矢印付きの破線にて示した。また、L 内の距離 $D_{feature}$ を任意の $x, y(\in H)$ に対し

$$D_{feature}(F(x),F(y)) \le D(x,y) \quad (8)$$

をみたすように定義する、つまり H での距離よりも L での対応する距離のほうが常に小さいように $D_{feature}$ を設定すれば、L 上で $D_{feature}(F(q),F(x)) \le \theta$ を満たす x の集合 $I_{feature}$ は、H 上で $D(q,x) \le \theta$ を満たす x の集合 I と、$I_{feature} \supseteq I$ の関係にある [1]。要するに L 上の特徴量を用いた近接点探索の結果に取りこぼしが起きないように特徴量関数 $D_{feature}$ を定義する。さらに、L 内の点を何らかの方法でクラスタリングしておくことにより、図中右側の長方形 c_1 から c_5 でしめしたようなクラスタが得られているものとする。

以上の準備の上で、探索は以下のように行う

- **Step 1A:**低次元空間でのクラスタ比較による除外 探索はまず L 内で行う。はじめに q とクラスタの低次元空間での位置関係比較にのみにより、クラスタ c_1 のような L 上でクラスタ全体が θ から外れるものが除外される。除外されたクラスタ内のすべての点は式 (8) を満たすから探索対象点となることはない。
- **Step 1B:**低次元空間での点比較による除外 次に Step 1A で除外できなかった、クラスタ c_2, c_3, c_4, c_5 のような場合について、L 内でクエリー点とクラスタ内の各点の間で距離比較を行い、$D_{feature}(F(q),F(x)) < \theta$ となる 点 x をすべて除外する。クラスタ c_2 のような q からの距離が θ 以内のものを含んでいないものは、ここで除外される。
- **Step 2:**高次元空間での除外 前 Step までの L での探索の結果は H 上では $D(q,x) > \theta$ となる点 (例えば図中点 a) を含むから、結果に対し、改めて H 上での比較を行い、真の正解を得る必要がある。そこで、前 Step までに除外されなかった点について、H 上で距離比較を行い、探索対象点を決定する。

以上の探索手法は、距離計算が高コストな H での探索 (Step2) に先立ち、あらかじめ計算した、式 (8) をみたす低次元な特徴量で探索対象点の絞り込みをし (Step 1A,1B)、全体としての探索コストを軽減する手法である。低次元空間である L においては、H に比べ容易に木構造やハッシュ構造などのクラスタリング手法が適用可能なので、L 上での絞り込みは、クラスタの包含関係を利用して絞り込む前段 (Step 1A) とその後の後段 (Step 1B) に分かれている。

3.2 疑似クラスタリングの導入

従来手法では絞り込みに利用した縮退特徴量空間ではクラスタリング手法が使われていたものの、その後の高次元空間での探索には使われていなかった。そこでわれわれは、高次元空間での探索にもクラスタ手法を導入する以下の手法を新たに提案する。

[提案手法]

H のクラスタリングを、H 上で直接行うのではなく、より低次元の空間である L でのクラスタリングの結果を用いて行う。従来手法において、すでに L 上でのクラス

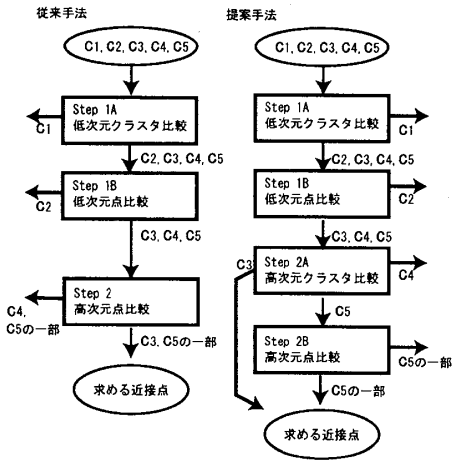

図 2: 従来手法と提案手法の処理の流れ

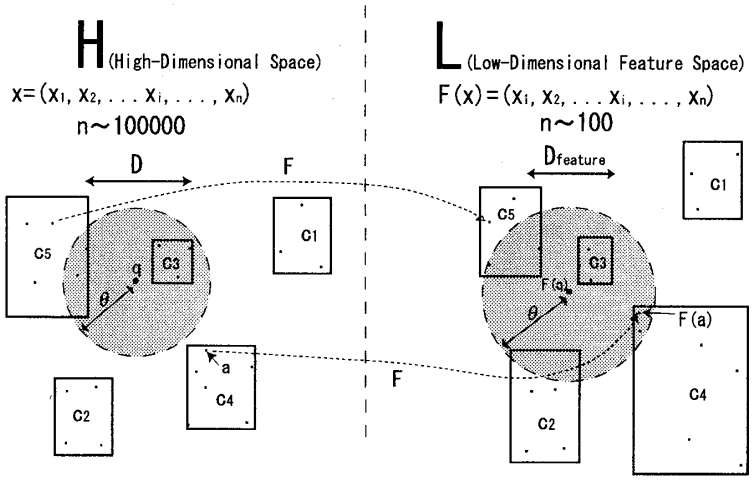

図 3: クラスタの包含関係

タリングはなされており、H 上のどの点が L 上のどの点に移されたかは既知であるから、図 3 右側の c_1 から c_5 に対応するものとして、H 上でも c_1 から c_5 のクラスタを構成できる。高次元空間でのクラスタとして利用するために、各クラスタについて、高次元空間での包囲図形を求めておくものとする。

以上の条件を追加した上で、前項の従来手法の Step2 を以下で置き換える。

- **Step 2A:**疑似高次元クラスタによる除外 H 上での q と各クラスタの位置関係により、以下のいずれかに分類する。以下において C はクラスタ内に含まれるデータ点の集合である。

  (a) $\forall x \in C, D(q,x) \leq \theta$: H 上においてクラスタ内の点がすべて θ 以内に含まれる場合は、q とクラスタとの位置関係比較でクラスタをまるごと含まれるとでき、クラスタ内の各点につき個別に H 上での距離を計算する必要はない。図中ではクラスタ c_3 がそれに該当する。

  (b) $\forall x \in C, D(q,x) > \theta$: 図中のクラスタ c_4 のように、L 上においては近接点とされたが、H 上においてクラスタ内の点がすべて θ より外にある場合、H 上で近接点とされた各点について個別に H 上の距離を計算する必要なく、すべて除外できる。

  (c) (a),(b) 以外: 図中のクラスタ c_5 のように、クラスタ内に H 上でも θ 以内のものと外れるものが混在する場合。

- **step 2B:**高次元空間での点間比較による除外前ステップ (c) に分類されたクラスタに属する点について q とクラスタの位置関係比較では判断できないため、クラスタ内の各点すべてについて H 上の距離 $D(q,x)$ を計算し θ と比較する。

3.3 コストの見積もり

従来手法と提案手法のコストを大まかに見積もって比較する。

数万次元以上の高次元空間での距離計算のコストに比べると、数十次元程度の低次元空間での距離計算や、クラスタの包含判定のコストはずっと小さく、手法全体に占める低次元空間での処理 (Step 1A,1B) のコストは通常たかだか数パーセントにすぎないので無視できる。

従って、従来手法の探索時間 $t_{previous}$ は Step 2 の探索時間 t_2 にほぼ等しい。いま、Step 2 における (a) に相当するクラスタが N_a 個、(b) が N_b 個、(c) が N_c 個あるとし、これらクラスタ内の点で Step 1B までで除外されなかった点がクラスタあたり平均 n_c 個あるものし、H での距離計算時間は 1 回あたり t_p とすると、$t_{previous}$ は

$$t_{previous} \approx t_2 = (N_a + N_b + N_c) n_c t_p \qquad (9)$$

となる。

提案手法の探索時間 $t_{proposed}$ についても同様に、$t_{proposed} \approx t_{2A} + t_{2B}$ であるから、H 上でのクラスタ包含判定に t_c の時間が掛かるとし、Step 2B での被探索点数のクラスタあたり平均を n_s とすると

$$t_{proposed} \approx (N_a + N_b + N_c) t_c + N_c n_s t_p \qquad (10)$$

となる。

ここで通常、高次元空間では距離計算のコストが巨大であるため、ともに 1 回の距離計算を含むクラスタの探索範囲内包含関係の判定と点の同判定ではほとんど差を生じず、

$$t_p \approx t_c \qquad (11)$$

であるから、探索時間比 $r := \frac{t_{proposed}}{t_{previous}}$ は

$$r = \frac{1}{n_c} + \frac{N_c n_s}{(N_a + N_b + N_c) n_c} \qquad (12)$$

となる。

クラスタリング手法依存であるが、通常 $n_c >> 1$ であるから第 1 項は十分小さいので、もし「N_a, N_b に比べて N_c が小さい」という条件が成立すれば提案手法による探索時間の短縮が望める。次節では同一映像断片探索に本手法を適用し、その有効性を検証する。

4. 実験

実験には 2002 年 10 月 21 日の NHK News 7 の映像 30 分 (352x240 pixels, 29.97 frames/sec., 53946 frames.)

を用いた。この映像を構成するフレーム画像から輝度情報のみを取り出し、輝度値により正規化した画像の集合を探索対象とした。距離関数は画像間距離である。特徴量関数 F としては、16 次元の正規化輝度ヒストグラムを抽出するものを用いた。特徴量空間での距離はヒストグラム間の L1 距離と定数値 d_{const} の積とした。この特徴量と距離では、厳密には 式 (8) を満たさないが、近接点探索の際に探索閾値 θ に応じて適切に d_{const} を設定することで、ほぼ取りこぼしがない探索が実現できることを確認している [3]。

特徴量空間でのクラスタ化は以下のように行った。SR-Tree [2] により木構造インデックスを構築した。SR-Tree は多次元空間に適用可能な木構造インデックスである。本実験で用いた SR-Tree の実装では、木構造の各ノードまたはリーフのメモリ上のサイズであるブロックサイズを固定的に与え、VAMsplit R-tree の構築アルゴリズムを使って、木構造を構築することができる。本実験ではこの機能を用い、得られた木構造の末端である各リーフノードを本手法におけるクラスタと見なした。VAMsplit R-tree の構築アルゴリズムは、可能な限りいっぱいに各ノードを充填しようとする。各点のデータサイズは一定 (浮動小数点数 16 個分のデータ領域+その点を示すラベルが占める数バイトの領域) であるから、対象となる点数が十分大きければ、与えたブロックサイズによりクラスタ (リーフノード) 内の点の数はほぼ等しく決定される。つまり、本構築法により、クラスタ内の点数をほぼ固定的に与えてクラスタリングを行っていることになる。

正規化画像空間 (高次元空間) におけるクラスタは、各クラスタ内に含まれる点をすべて含む半径最小の球をそのクラスタの包囲図形として用いた。正規化画像空間でクエリー点とクラスタとの位置関係を得るには、正規化画像空間での距離計算 1 回とそれに比べると無視できるほどわずかな計算量の和の計算や比較のみ で済むので、式 (11) が成立する。

さて、クエリー点としては、実際の探索と同様の結果が得られるように、探索対象と同一フォーマットの別の日 (2002 年 10 月 14 日) の News 7 の映像から 10 フレームごとに 1 フレーム抜き出すことにより生成した、5395 個の画像を用いた。クエリーによる偏りが出ないように、これら画像で順次に探索を画像枚数回行い、探索全体での N_a, N_b, N_c 等を測定した。

すでに述べたようにクラスタの構築に VAMsplit R-tree のアルゴリズムを用いているので、一つのクラスタ内に含まれる点の数 n_c は、固定とみなせる。構築時に与えたブロックサイズに従って、n_c が決定される。本手法の性能を左右する N_a, N_b, N_c 等は n_c により変化すると予想される。そこで、n_c を変化させて得られた、N_a, N_b, N_c 等を図 4 に示す。N_a,N_b,N_c,$N_c n_s$ の和がグラフの高さになっており、式 (11) と (10) からも分かるように、H 上での距離比較の回数 $\frac{t_{proposed}}{t_p}$ を表している。

n_c が大きくなるにつれ、N_a, N_b が小さくなり、$N_c n_s$ が大きくなる傾向が見られる。結果は、直感的な予想とも一致しており、$n_c = 19$ で $\frac{t_{proposed}}{t_p} = 6425$ の最小を示した。

一方従来手法での距離比較の回数 $\frac{t_{previous}}{t_p}$ は 35070 であり、図 4 では点線に相当する。提案手法では従来手法に比べ比較計算回数が約 5.5 分の 1 となり、大幅な高次元距離比較計算削減効果が確認できた。

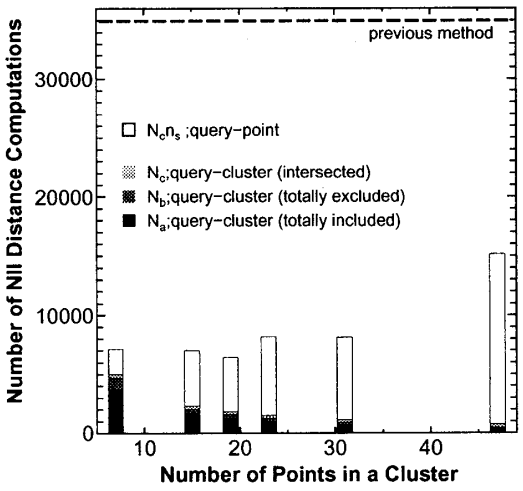

図 4: クラスタ内画像数と画像間距離計算数

5. まとめ

同一映像断片探索は高次元での近接点探索問題と等価であり、次元圧縮をした特徴量空間で空間のクラスタリングをすることによる高次元空間内近接点探索の高速化の手法は、実際の放送映像で大きな高速化効果をもたらすことが確認できた。ただし、本手法の適用のためには、あらかじめ特徴量を計算し特徴量空間でクラスタ化を行い、さらにそのクラスタ化の結果に基づいて、元の高次元空間でのクラスタ包囲図形の計算を行っておく必要がある。したがって、本手法による高速化はこれら前処理にかかる余分な処理とトレードオフとなる。これら前処理は探索対象となるデータが静的であれば、一度行っておけば済むため、全体としての効率は探索対象データの更新の頻度に主として依存する。同一映像断片探索のような、対象となるデータがほとんど静的で巨大な場合、前処理のコストに比べて高速化の効果の方が占有的となるため、本手法は有効である。

参考文献

[1] Carlo Zaniolo, Christos Faloutsos, V. S. Subrahmanian, Stefand Ceri, Richard T. Snodgrass, Roberto Zicari, “Advanced Database Systems”,Part IV, Morgan Kaufmann Publishers, Inc., 1997.

[2] 片山 紀生, 佐藤 真一, “SR-Tree: 高次元データに対する最近接点検索のためのインデックス構造の提案”, 電子情報通信学会論文誌,vol. J80-D-I 1997-8,pp.703-717,1997

[3] 山岸 史典, 佐藤 真一, 浜田 喬 “大規模映像アーカイブのための映像断片照合の高速化”, 第 1 回 情報科学技術フォーラム講演論文集, LI-17, 2002.

[4] 山岸 史典, 佐藤 真一, “同一映像断片探索に基づくニュース映像ブラウザの実装”, 電子情報通信学会技術研究報告, PRMU, Dec. 2003.