4B-2

大規模多次元特徴ベクトルのハッシュに基づく検索法の性能比較

村林　昇†　　吉田　健一†

筑波大学大学院　ビジネス科学研究科†

1. はじめに

大規模な多次元特徴ベクトルの高速検索手法として，我々は先の研究[1]で Tiny LSH 法を提案した．Tiny LSH 法は，検索対象である入力データの特徴ベクトルを補正関数に基づくビットシフト処理と Direct-mapped cache を組み合わせた検索法であり，従来の Locality Sensitive Hashing(以下 LSH)[2]と同様に類似データの検索を，少ないハッシュ関数とメモリ空間で高速に行うことができる．本研究では，Tiny LSH 法とチェイン法および従来の LSH 法との検索性能の比較を，多次元データベースにおける近似検索方法について考察した．

2. Tiny LSH 法に基づく高速検索手法

2.1　RGB 画像特徴量と Tiny LSH 処理

文献[3]は，RGB 色区間に基づいた画像特徴量を用いたビデオ検索の研究で検索時間は長いが検索性能が良い結果を報告している．本研究ではその手法と同様のアプローチをとった．図 1 に RGB 色空間に基づいた画像特徴量の処理方法を示す．

図 2 に Tiny LSH 処理を示す．画面が暗いフレームでは特徴データが 0 となる傾向があるのでそれを改善するためにデータの量子化をビットシフト処理によって変えている．

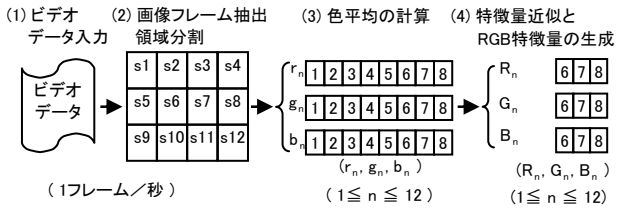

図 1　RGB 画像特徴量の処理方法

補正関数　: Ft(Rn, Gn, Bn) (1 ≦ n ≦ 12)

```
1: Cn = ( Rn >> 1 + Gn >> 1 + Bn >> 1) >> 1;
2: if ( Cn == 0 );
3:    Cn = ( Rn >> 1 + Gn >> 1 + Bn >> 1 );
4: if ( Cn == 0 );
5:    Cn = Rn + Gn + Bn ;
6: Return Cn ;
```

図 2　　Tiny LSH 処理

2.3 Direct-mapped cache に基づいた高速検索

提案手法は，RGB 特徴量に対して，Tiny LSH 法の概念と文献[4]で用いた高速データ検出の手法を組み合わせて検索処理を高速化した．

図 3 に，RGB 特徴量に対する Direct-mapped cache と原ビデオデータベースからなるハッシュテーブルのデータ構造を示す．検索のためのビデオ画像のフレーム間距離は図 1(4)で生成した RGB 特徴量を用いてユークリッド距離を計算する．

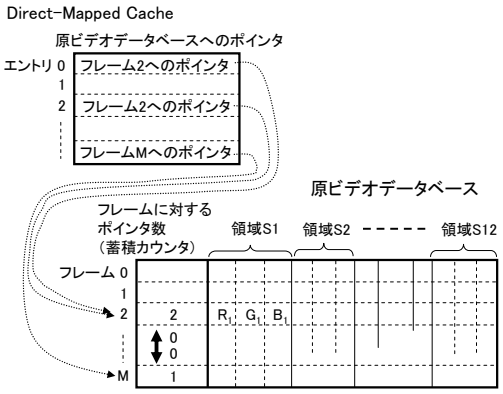

図 3　画像特徴量の Direct-mapped cache

3. 実験

3.1　多次元特徴ベクトル

性能比較を行うために、多次元特徴ベクトルとして実際のビデオデータを用いた．ビデオデータは，The Open Video Project の Web サイト[5]から総時間 109.5 時間分，380 本のビデオコンテンツをダウンロードした．ダウンロードしたビデオコンテンツから図 4 に示すような，画像サイズとビットレートを変えた Re-scaled ビデオを生成しテストデータとして用いた．

Performance Comparison of the hash-based date retrieval methods for large-scale multi-dimensional feature vectors.
† Noboru Murabayashi
† Kenichi Yoshida
Graduate School of Business Sciences, University of Tsukuba
(†) Otsuka 3--29--1, Bunkyo-ku, Tokyo, 112--0012 , Japan

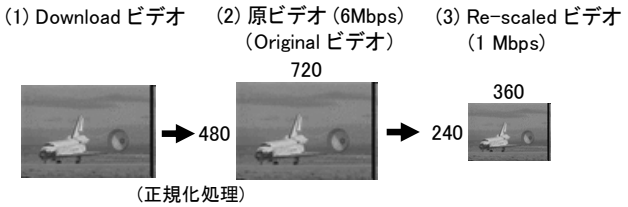

図 4　テストデータの生成方法

3.2　実験結果

(1)Tiny LSH 法とチェイン法の検索性能比較

従来手法と提案手法である Direct-mapped cache の比較で主たるポイントは検索時間である．ビデオのデータ量が変わると類似フレームの量が変わりハッシュ値の衝突頻度が変わる．そのようなことから，検索に用いるビデオのデータ量を 3.4 時間～109.5 時間の間で変化させて実験を行い，提案手法と従来手法の性能比較を行った．図 5 に示す正解率のデータから，136 年分のデータ量では，チェイン法の場合は約 94%，提案する Tiny LSH 法の場合は約 64%になると推定できる．チェイン法では，ハッシュ値の衝突が起きてもハッシュテーブルのリンク長が増加することでデータは残るので検索できるが，Tiny LSH 法では図 3 で説明したハッシュテーブルにおけるポインタの上書きによって残るポインタの確率が低下するために正解率が低下すると考えられる．チェイン法は正解率の大きな低下はないが図 6 の結果から平均検索時間は実用レベルの性能ではないことが分かる．

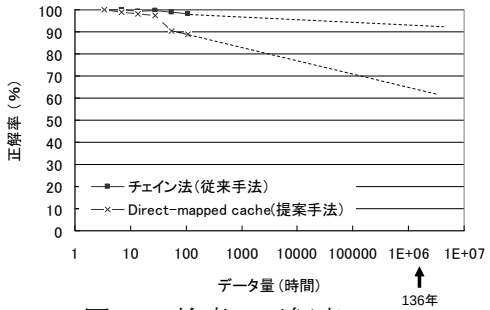

図 5　検索の正解率

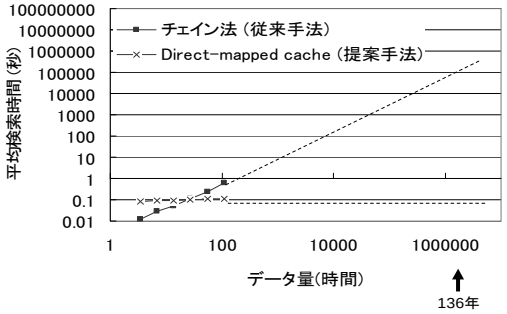

図 6　検索時間の結果

(2) LSH との類似フレーム検索性能比較

類似フレームの検索性能について，Tiny LSH 法と従来の LSH を比較する実験を行った．図 7 に処理時間（一つのハッシュ値一致検索の時間＋データ蓄積時間）を示す．従来の LSH は Tiny LSH 法よりも処理時間が長く，ハッシュ関数の数が増加すると処理時間も増加する．一方，Tiny LSH 法は，測定誤差の範囲で同じ処理時間であった．

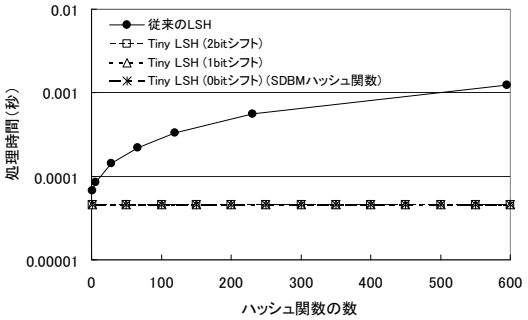

図 7　ハッシュ値の一致検索処理時間

4.　まとめと今後の課題

大規模な多次元データベースにおける近似検索方法として Tiny LSH 法と従来手法であるチェイン法，LSH の検索性能の比較検討を行った．実験により Tiny LSH 法は従来手法よりも優れていることを示した．今回よりも大規模，改変の種類を多くしたデータベースにおける近似検索実験の確認が今後の課題である．

参考文献

[1]K.Yoshida and N.Murabayashi,"Tiny LSH for Content-based Copied Video Detection",Proc. of SAINT2008, IEEE CS, pp.89-95, 2008

[2]M. Datar, P. Indyk, N. Immorlica and V. Mirrokni," Locality-sensitive hashing using stable distributions",Nearest Neighbor Methods in Learning and Vision: Theory and Practice, MIT Press, 2006

[3]長坂 晃朗,宮武 孝文,"時系列フレーム特徴の圧縮符号化に基づく映像シーンの高速分類手法",電子情報通信学会論文誌 D-11,J81-D-11, No.8, pp. 1831-1837, 1998

[4]Kenichi Yoshida, Fuminori Adachi,Takashi Washio, Hiroshi Motoda, Teruaki Homma, Akihiro Nakashima, Hiromitsu Fujikawa and Katsuyuki Yamazaki,"Density-based spam detector",KDD2004, pp.486-493, 2004

[5] http://www.open-video.org/