クラスタリングに基づく情報の検索と視覚化

堀田, 政二

https://doi.org/10.15017/1398256

第8章

次元圧縮とクラスタリングに基づく画像の近似kNN検索

8.1 まえがき

本章では，画像の近似kNN検索の手法として，主成分分析による特徴ベクトルの次元圧縮とクラスタリングを組合せた方法を提案する[42]．前章では，特徴ベクトルの次元削減とクラスタリングに基づくフィルタリングを用いて画像のkNN検索を高速化する手法を提案した．この手法はフィルタリングによって計算量を削減してはいるが，正しい検索結果を出力するのを保証するために，高速化率はあまり高くない．しかし，実際にはk個すべてが正しい結果を出力する保証は必ずしも必要ではない．なぜなら最初に入力されるクエリは所望の画像から遠いことが多く，1回めの出力画像k枚のなかに所望のものが見つかるのは稀であり，通常はクエリを変えながらインタラクティブに検索が進められる．そのような場合，個々のkNN検索ではk個ともすべて正しい結果が出力されなくても，クエリの移動によってカバーすることができると考えられ，むしろ個々のkNN検索の応答の速さが重要になると思われる．最近ではそのような近似kNN検索の研究が盛んに行われている[18]．

本章では次元圧縮とクラスタリングに基づく近似kNN検索法を提案する．検索結果は必ずしも正確でなくてもよいとはいえ，近似精度は高い方が望ましく，検索結果の近似度を上げるには次元圧縮後の距離の変化が小さい方がよい．前章ではカラーヒストグラムの2次形式距離の標準化に基づいて次元を削減したが，この方法では距離の保存度はあまり高くない．周知のように距離の保存度が最も高い線形次元圧縮法は主成分分析である．そこで本章では主成分分析を用いることにする．また，前節と同様に次元圧縮に加えて，クラスタリングも用いて探索範囲を限定することによって検索の計算量を更に減らす．

8.2 特徴ベクトルの次元圧縮

本提案手法では画像の特徴ベクトルとしてカラーヒストグラムを，次元圧縮法としては主成分分析を用いる．画像間の距離を単純に n 色のカラーヒストグラム $\boldsymbol{h} = [h_1, ... h_n]^T$ のユークリッド距離とすれば，ヒストグラムに直接主成分分析を適用することができ，特徴ベクトル，すなわちヒストグラムの次元を圧縮することができる．しかし，この単純な距離は視覚的な色の非類似度からの歪みが大きく，あまり良好な検索結果は得られない．そこで画像間の距離を2次形式距離で測ることにする．ただし，前章の式 (7.2) の2次形式距離のままでは主成分分析を適用できない．そこで前章で述べたように，カラーヒストグラム $\boldsymbol{h}$ を前章の式 (7.3) で $\boldsymbol{x}$ に標準化すれば，画像間の2次形式距離は $\boldsymbol{x}$ のユークリッド距離となる．そこで $\boldsymbol{x}$ に主成分分析を適用すれば次元を圧縮することができる．

なお Hafner ら [14] の次元削減法は，2次形式を標準化した段階 $D(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j) = \sum_{r=1}^{n}(x_{ir} - x_{jr})^2$ で，これを l 項までの部分和 $\sum_{r=1}^{l}(x_{ir} - x_{jr})^2$ で近似する方法である．本章の次元圧縮法では2次形式の標準化の後で更に主成分分析する必要があり，Hafner ら [14] の方法よりも一見手間が増えそうであるが，標準化の変換行列も主成分分析の射影行列もデータベースから予め計算でき，2つの行列を1つにまとめれば，本方法の変換の手間は Hafner らの方法と同じである．

8.2.1 次元圧縮による距離の変化

本章では上記のように2次形式の標準化の後で主成分分析を行い，次元を圧縮する (以後これを SP-QFD[1] と呼ぶ) が，画像間の距離がこの次元圧縮によってどの程度変化するか実験した．比較のため，カラーヒストグラム $\boldsymbol{h}$ のユークリッド距離の主成分分析による方法 (以後 P-ED[2] と呼ぶ) についても調べた．前節でも述べたように，この距離は視覚的な類似度からの歪みが大きく，あまり使われないが，比較のためにあえてここで調べた．また，Hafner ら [14] の方法，すなわち2次形式の標準化だけによる方法 (以後 S-QFD[3] と呼ぶ) についても調べた．P-ED と S-QFD および SP-QFD とでは元の距離が異なるため，そのまま比較することはできない．そこで，P-ED と SP-QFD では次元圧縮後の距離と元の距離との比を，S-QFD では次元削減後の距離

[1] dimensionality reduction by Standardization and Principal component analysis for Quadratic Form Distance の略.

[2] dimensionality reduction by Principal component analysis for Euclid Distance の略.

[3] dimensionality reduction by Standardization for Quadratic Form Distance の略.

と元の距離との比を調べることにする．3つの方法すべてにおいて，この比は1より小さくなるが，1に近いほど距離の近似精度が高いことになる．

実験にはサイズ 150×150 の風景写真500枚を用いた．色数 n は64とした．実験の結果を図8.1に示す．横軸は低次元数である．実線は本提案法SP-QFDであり，破線はP-EDで，点線はHafnerらの方法S-QFDである．本方法が最も近似の精度が高く，$l = 15$ でほぼ1になった．SP-QFDがP-EDよりも精度が高いということは，2次形式距離がユークリッド距離よりも視覚心理的に優れているのみならず，次元の圧縮性能も優れていることを示しており，2重の意味で検索に適しているといえる．

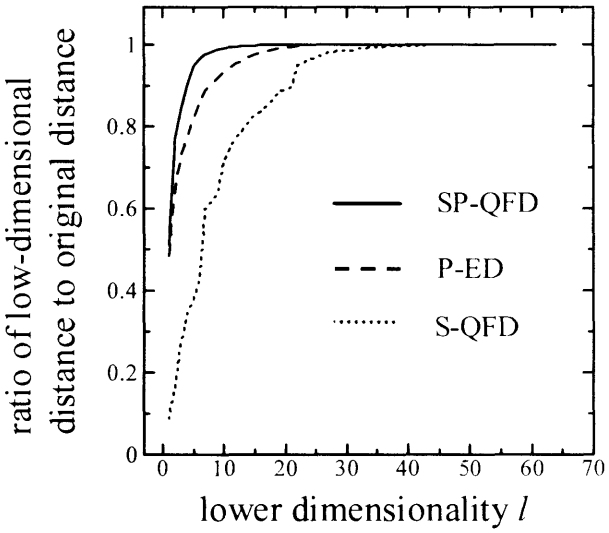

図 8.1: 低次元距離と元の距離との比

8.3 次元圧縮とクラスタリングによる近似 kNN 検索

以上のように次元圧縮法SP-QFDを用いれば $l = 20$ 程度の低次元でほとんど正しい検索結果が得られる．元の次元が $n = 64$ であるから，この次元圧縮によって計算量が約1/3に削減できる．本章ではこの次元圧縮に加えてクラスタリングも用いて計算量を更に削減する．すなわち，データベースの画像を予めクラスタリングしておき，クエリが入力されたらまず最初にクエリに近いクラスタを何個か選び，それらのクラスタに含まれる画像のなかからクエリに近いものを k 個選んで出力することにする．これが本章で提案する検索法であり，Hafnerら[14]の方法が2次形式の標準

化による次元圧縮だけを用いるのに対し，本方法では更にそれに引き続き主成分分析とクラスタリングを行なう．簡略に記すと，Hafner らの方法は標準化のみの厳密検索，前章では標準化＋クラスタリングの厳密検索，本提案法は標準化＋主成分分析＋クラスタリングの近似検索である．なお厳密と近似の違いは，正しい kNN の出力が保証されるか，あるいは必ずしも保証されないかの違いである．

このときの計算量について簡単に検討する．まず，本方法ではクラスタの大きさは揃っている方がよい．なぜなら，クラスタ i の画像数を Q_i，データベースの総画像数を Q とすると，クラスタ i が選ばれる確率は Q_i/Q にほぼ比例し，クラスタ i のなかの検索の計算量は Q_i に比例するから，平均計算量は $\sum_i Q_i^2/Q$ に比例するので，クラスタの数を N とすると

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} Q_i^2 \\ \text{subj.to} \quad & \sum_{i=1}^{N} Q_i = Q \end{aligned} \tag{8.1}$$

の解 Q_i のときに計算量が最小となり，式 (8.1) の解はすべての Q_i が同じ ($Q_i = Q/N$) ときであるからである．

次にクラスタの数 N について考える．まずクエリに近いクラスタを選ぶのに N に比例する計算量を要し，クラスタの大きさが揃っているとすると各クラスタの大きさは Q/N となるので，M 個のクラスタを選ぶとすると選ばれるデータ数は MQ/N となり，そのなかから k 個選ぶ計算量は MQ/N に比例するので，総計算量は $N + MQ/N$ に比例する．これが最小になる N は $\sqrt{MQ}$ となるが，近似度があまり小さくならないようにここではこの 1/3 程度のクラスタ数とすることにする．

なお，クラスタリングは予め 1 回行うだけなので元の次元で行ってもよいが，ここではクラスタリングの時間も短縮するために低次元で行うことにする．

8.3.1 低次元でのクラスタリング

ここではクラスタリングに k 平均法を使う．上記のようにクラスタの大きさは揃っている方がよく，k 平均法をそのように各クラスタの大きさをできるだけ揃えるように修正することもできるが，ここでは単純に，初期値を変えて何回か k 平均法を行い，クラスタの大きさの分散が最も小さい解を採用することにする．

SP-QFD でデータベース画像の次元を圧縮し，低次元で k 平均法を適用する．この低次元でのクラスタリング結果が元の次元での結果からどの程度ずれるかを次のようなエントロピーで評価した．元の次元でのクラスタ i に属す画像のなかで低次元

でのクラスタ j に含まれるものの個数を n_{ij} とする．$p_{ij} = n_{ij} / \sum_{i=1}^{N} n_{ij}$ とするとクラスタ j のエントロピーは

$$H_j = -\sum_{i=1}^{N} p_{ij} \ln p_{ij} \tag{8.2}$$

であり，全エントロピーは

$$H = \sum_{j=1}^{N} S_j H_j \tag{8.3}$$

である．ここで S_j は元の次元でのクラスタ j に含まれる画像の個数である．この H が小さいほど，低次元でのクラスタリングは元の次元での結果に近いことになる．

クラスタ数を10とした場合のエントロピー H を図8.2に示す．データとして8.2.1節と同じ風景写真500枚を用いた．横軸は低次元数である．比較のためP-EDとS-QFDの結果も示した．この図の H は初期値を変えて100回実験した平均値である．提案法SP-QFDでは $l = 20$ 辺りでエントロピーがほぼ0になり，元の次元でのクラスタリングとほぼ同じ結果を与えることがわかる．クラスタリングの計算時間は $l = 20$ のSP-QFDで0.46秒であり，元の次元での時間1.16秒よりも短い．次節での検索の実験ではこの $l = 20$ のSP-QFDでのクラスタリング結果を使うことにする．

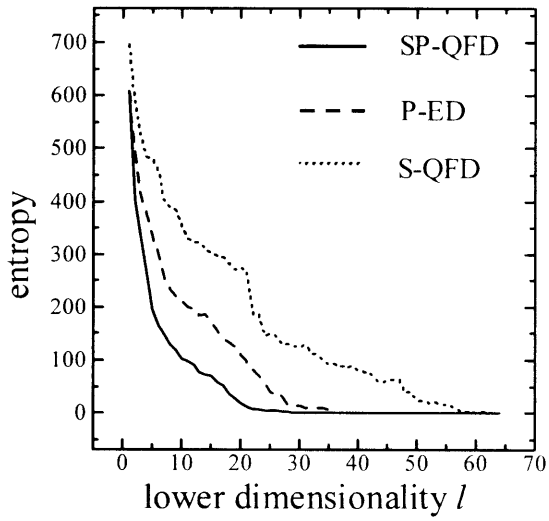

図 8.2: 低次元クラスタリングのエントロピー

8.3.2 近似kNN検索

データベースを上記のように予めクラスタリングして，各画像がどのクラスタに属すかという情報と各クラスタの代表ベクトルを保存しておく．前節のようにここでは20次元でクラスタリングするのでメモリされる代表ベクトルも20次元である．クエリ画像が入力されたら，まず最初にクエリに近いクラスタを何個か選ぶ．このときクエリも20次元に圧縮して代表ベクトルとの距離を計算する．次に，選ばれたクラスタに含まれる画像全部のなかからクエリに近いものをk個選ぶ．このときもクエリと各画像との距離は低次元で計算する．

これがここで提案する近似kNN検索法であるが，元の次元で選択されるk個のうちの何個かは漏れる可能性がある．そこで前節の風景写真500枚について$k=10$として実験し，元の次元での10個のうち何個が本近似検索の結果に含まれるかを調べた結果を図8.3に示す．クラスタの数は10個である．横軸はクエリと各画像との距離を計算するときの次元数である．選択するクラスタの数を1個から10個まで変えてみたが，4個以上ではほとんど同じであったので図8.3には4個までを示す．例えば$l=20$ではクラスタ数1のとき7.91，2個のとき9.5，3個のとき9.86，4個で9.97となり，$l=20$でクラスタ数2個で元の次元のkNNとほぼ同じ結果が得られる．次に検索時間を図8.4に示す．クラスタ数2個で$l=20$では0.0013秒である．元の次元でのkNNは0.013秒であり，本方法はこれの10倍速い．理想的には計算時間は，クラスタ選択により2/10，次元圧縮により20/64，合わせて40/640=0.0625倍に短縮されるはずであるが，実際にはこれより少し遅くなっている．なお，同じデータで$k=10, l=20$のときのHafnerら[14]の方法は0.0044秒かかり，前章での方法では0.0038秒であった．本方法はこれの約3倍速い．すなわち，近似検索にすることによって，複雑なフィルタリングアルゴリズムを使うことなしに，ほぼ同じ検索結果を与えるのに約3倍高速化できた．

なお本方法は，次元削減によるフィルタリングを用いることによりさらに高速化することができる．高速化できる箇所は，クエリに近いクラスタを選ぶところと選ばれたクラスタに含まれるデータのなかからクエリに近いものをk個選び出すところの2箇所である．この2箇所で7.2節のアルゴリズムkNN-OFDRを用いることにより本方法の検索時間は0.0006秒になり，0.0013秒よりも更に約2倍高速化できた．

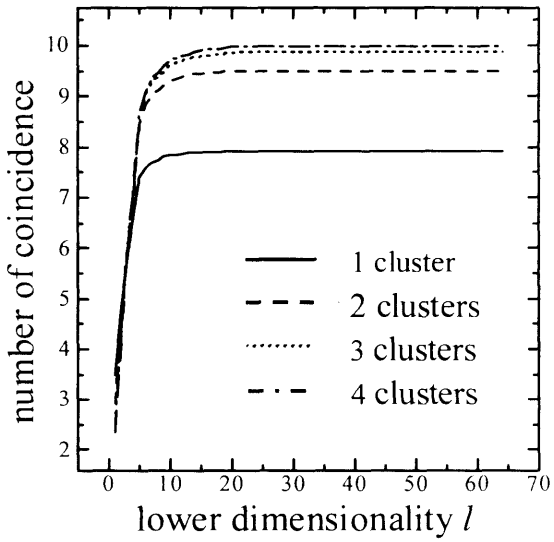

図 8.3: 元の次元での kNN との一致個数

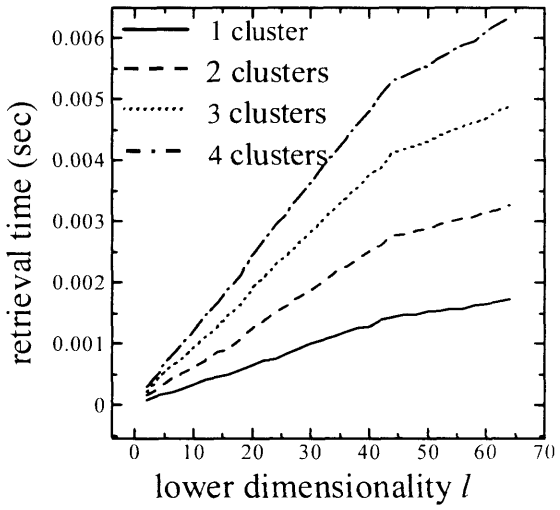

図 8.4: 検索時間

8.3.3 色数 n による違いについて

以上，本実験では色数 n が 64 の場合について調べた．一般にヒストグラムの色数を増やすと情報の冗長度が増し，したがって，次元圧縮率を高くすることができるようになる．それは例えば図 8.5 からもわかる．図 8.5 は，提案法 SP-QFD について図 8.2 でも示したエントロピーを色数が 64 の場合 (これは図 8.2 と同じ) と 512 の場合について示したものである．これから，$n = 64$ ではほぼ $l = 20$ 次元に圧縮できていた (圧縮率ほぼ 1/3) のに対し，$n = 512$ では約 64 次元まで圧縮する (圧縮率 1/8) ことができ，検索の高速化率も高くなる．このように本方法の有効性は色数が多くなるほど高くなる．

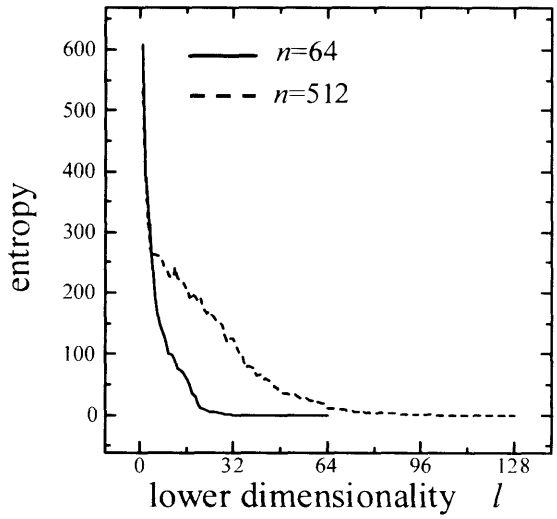

図 8.5: 低次元クラスタリングのエントロピー

8.4 むすび

画像の近似 kNN 検索法として，2 次形式の標準化と主成分分析による次元圧縮，およびクラスタリングに基づく方法を提案し，500 枚の画像データで実験した結果，元の次元での全探索とほぼ同じ結果を出力するのに約 10 倍，前章で提案したフィルタリング法と比べても約 3 倍高速化することができた．また，本方法にも次元削減によるフィルタリングを用いることにより，更に約 2 倍高速化できた．