ロバスト潜在意味解析に基づくビデオの類似検索

Similarity Search for Videos Based on Robust Latent Semantic Analysis

正会員 井上 光 平[†], 浦 浜 喜 一[†]

Kohei Inoue[†] and Kiichi Urahama[†]

Abstract A method retrieving videos is presented by utilizing vector quantization and latent semantic analysis. Each video is represented by a sequence of signatures through the vector quantization of frame datasets. Latent semantic analysis is then applied to the signature with a video matrix. We verified through experiments that dimensionality reduction in latent semantic analysis increases the speed and precision of retrieval. Making vector quantization more robust further improved the performance of similarity searches.

キーワード: ビデオ検索, 潜在意味解析, ロバストベクトル量子化

1. ま え が き

文書検索の手法として開発された潜在意味解析 (Latent Semantic Analysis: LSA)¹⁾ はマルチメディア検索にも応用され, キーワード検索に加え, カラーヒストグラム検索²⁾ や適合性フィードバックに基づく類似検索³⁾ などに使われている. 筆者ら⁴⁾ も前に, ビデオの類似検索への応用として, フレーム集合の階層クラスタリングに基づいてビデオを記号列に変換して, LSA によってビデオ間の類似度を求める手法を提案した. しかしその手法では, どの代表フレームからも遠いようなフレームも残らずどれかの記号に置換えられ, そのような例外値によって検索性能が落ちる可能性がある. 例えば, 放送ビデオ中の CM やニュース映像でのフラッシュ, 映画の黒画面による切れ目などのように, ビデオの主題とは無関係な短いショットが混在することが多く, ビデオの類似検索では, そのような例外ショットは除くのが妥当と考えられる.

そこで本論文では, クラスタリングを k 平均法に変えてロバスト化するとともに, ビデオを記号列に変換するベクトル量子化 (Vector Quantization: VQ) もロバスト化して, 例外値の影響を軽減した LSA によるビデオの類似検索法を提案する.

2. 潜在意味解析とベクトル量子化

ここで提案する手法の基となる LSA と VQ を概説して

おく. 文書が n 個あり, 単語が m 個あるとし, 単語 i が文書 j に出現する頻度を d_{ij} とする. 文書 j はベクトル $d_j = [d_{1j}, \dots, d_{mj}]^T$ で表され, 通常の実験では文書 j と j' の類似度は d_j と $d_{j'}$ のコサインで測られるが, LSA¹⁾ では低次元でのコサインを類似度として用いる. すなわち, $m \times n$ データ行列 $D = [d_{ij}]$ を $U\Sigma V^T$ と特異値分解し, k 行, k 列だけ残して階数 k の近似行列 $U_k \Sigma_k V_k^T$ とし, ベクトル d_j を $\hat{d}_j = d_j^T U_k \Sigma_k$ で $k (< m)$ 次元に射影して, $\cos \theta_{j,j'} = \hat{d}_j \hat{d}_{j'}^T / \|\hat{d}_j\| \|\hat{d}_{j'}\|$ を文書 j と j' の類似度とする. この手法の利点は, 階数を落とすことによってデータのノイズや冗長性および類似度の計算量が低減され, 検索精度と検索速度が向上することである.

次に VQ は, クラスタリングに基づくデータ符号化の手法であり, 検索の高速化に適用されている. n 個のデータからサンプリングしたトレーニングデータを $l (< n)$ 個のクラスタに分割し, 各クラスタの代表データをコードワードと呼び, l 個のコードワードの集合をコードブックと呼ぶ. n 個の各データは自分に最も近いコードワードのクラスタを表す記号 (通常はクラスタの番号) に符号化される.

3. ビデオの類似検索法

本論文で提案する方法では, VQ によってビデオを記号列に変換して, ビデオを文書, 記号を単語とみて LSA でビデオ間の類似度を計算し, クエリビデオと似ているビデオを検索する.

3.1 ビデオの記号列への変換

まず最初に, VQ によって各ビデオを記号列に変換する. 全データベースビデオから 10 フレームおきにサンプリングしたフレームの集合を VQ のトレーニングデータとし, それを k 平均法でクラスタリングする. ここでは, 各フレー

2004 年 5 月 14 日受付, 2004 年 9 月 15 日最終受付, 2004 年 10 月 7 日採録

[†]九州大学 大学院芸術工学研究院

(〒 815-8540 福岡市南区塩原 4-9-1, TEL 092-553-4512)

[†]Faculty of Design, Kyushu University

(4-9-1, Shiobaru, Minami-ku, Fukuoka-shi, 815-8540 Japan)

ムをRGB各2ビットの64色のカラーヒストグラムで表し、ヒストグラム間のユークリッド距離をフレーム間の非類似度とした。各クラスタの重心を代表フレームと呼び、VQのコードワードとする。次に、ビデオの各フレームを自分に最も近い代表フレームのクラスタの記号(番号)に置換える。このとき、フレームの時間的な連続性を利用して次のようにする。

変換法 1:

Step 0: 各クラスタの半径(そのクラスタの代表フレームから最も近い代表フレームまでの距離の1/2(図1の r))を計算する。

Step 1: ビデオの第1フレームと各代表フレームとの距離を計算し、最も近い代表フレームに置換える。

Step 2: 第2フレーム以降については、一つ前のフレームが置換えられた代表フレームとの距離を計算し、それがその代表フレームの半径以下(図1の Δ)であればそれに置換え、そうでない(図1の $\diamond$ や $\star$)なら他の全代表フレームとの距離を計算し、最も近い代表フレームに置換える。

しかしこの方法では、どの代表フレームからも遠いようなフレーム(図1の $\star$)も必ずどれかの記号に置換えられ、そのような例外値によって検索性能が落ちる可能性がある。そこで、最も近い代表フレームとの距離がその代表フレームのクラスタの半径の $\alpha(>1)$ 倍よりも大きいフレームは無視することにする。

そのためにまず、代表フレームを求める k 平均法をロバスト化する。すなわち k 平均法では、各代表点を自分に属すデータの重心に移動するのを繰り返すが、そのとき自分の半径の $\alpha(>1)$ 倍よりも遠いデータは自分に属さないとして無視する。またビデオを記号列に置換える手順(上記の変換法1)も次のように変える。

変換法 2:

Step 0: 各クラスタの半径(図2の r)を計算する。

Step 1: 第1フレームと各代表フレームとの距離を計算し、最も近い代表フレームに仮に置換える。

Step 2: 第2フレーム以降については、一つ前のフレームが置換えられた代表フレームとの距離を計算し、それがその代表フレームの半径以下(図2の Δ)であればそれに置換え、そうでない(図2の $\diamond$ や $\star$)なら他の全代表フレームとの距離を計算し、最も近い代表フレームに仮に置換える。

Step 3: 上で仮に置換えられた代表フレームとの距離がその半径の α 倍以内(図2の $\diamond$)なら、その置換えを確定し、 α 倍よりも大きい(図2の $\star$)ならそのフレームは無視する。

3.2 潜在意味解析による類似検索

以上によって、各ビデオは記号列に置換えられる。記号の数を m 、データベースビデオの数を n とする。記号 i のビデオ j での出現頻度を d_{ij} とすると、ビデオ j はベクトル $d_j = [d_{1j}, \dots, d_{mj}]^T$ で表され、データベース全体は $m \times n$ 行列 $D = [d_{ij}]$ で表される。LSAに倣って D を $U\Sigma V^T$ と特異値分解し、 k 行、 k 列だけ残して階数 k の近

図1 代表フレーム(●)の半径(r)
Radius (r) of representative frame(●).

図2 代表フレーム(●)の半径(r)と棄却半径(αr)
Radius (r) and rejected radius (αr) of representative frame(●).

似行列 $U_k \Sigma_k V_k^T$ とする。ベクトル d_j を $\hat{d}_j = d_j^T U_k \Sigma_k^p$ で $k(<m)$ 次元ベクトルに射影する(ここで p は可変な指数として導入した)。クエリビデオについても上記の変換法によって記号列に置換えて記号の出現頻度を計算し、ベクトル $q = [q_1, \dots, q_m]^T$ で表し、 $\hat{q} = q^T U_k \Sigma_k^p$ によって k 次元ベクトルに射影する。ビデオ間の類似度を射影後のコサイン: $\cos \theta_j = \hat{q} \hat{d}_j^T / \|\hat{q}\| \|\hat{d}_j\|$ で求め、この値が大きい順にビデオ j を出力する。コサインの計算では、オフラインで $\hat{d}_j / \|\hat{d}_j\|$ を計算しておき、 q が入力されたら $q^T U_k \Sigma_k^p / \|q^T U_k \Sigma_k^p\|$ を計算して $\hat{d}_j / \|\hat{d}_j\|$ との内積を計算する。この内積は k 次元なので元の m 次元よりも計算量が少ない。

なお、各ビデオの長さ(フレーム数)は異なっているかもしれない。 q_i や d_{ij} はビデオの長さに比例するが、コサインの計算のときに q と d_j のノルムで割るので単位長さ当たりの頻度に変換され、類似度は長さによらないからである。

4. 実験

フレームサイズ320画素×240ライン、RGB色空間、フレーム周波数29.97Hz、フレーム数300のビデオ40本(海辺、夕景、サッカー、雪景色の4グループ各10本ずつ)を用い、各ビデオをクエリとしたときに、同じグループのビデオが上位10本のうちに何本含まれるか調べた。クラスタ(記号)の数 m は、この例では10~30程度の範囲で結果はあまり変らなかったため、 $m=20$ とした。このことから逆に、これらのビデオでは、各カテゴリが約5個のクラスタで形成されていると推測される。すなわち m は一般に

図 3 検索時間
Computational time for retrieval.

図 5 ロバスト VQ でのグループの一致率
Retrieval rate with robust VQ.

図 4 上位 10 位中のグループの一致率
Retrieval rate of correct group.

は、各カテゴリのクラス数 $\times$ カテゴリ数を目安とすればよいと思われる。

LSA を使わない基礎的な手法として、元のベクトル q, d_j 間のヒストグラムインタセクションを類似度として検索したところ、上位 10 本中平均 7 個のビデオがクエリとグループが一致した。検索時間は 500 回当たり 0.92 秒であった。

LSA による検索では、射影行列 $U_k \Sigma_k^p$ の p を -1 から 3 まで変えて実験した結果、 $p = 2$ のときが最もよかったので $p = 2$ とした。 p が大きいほど固有値の比が強調され、ビデオのように比較的ノイズが大きいデータでは $p = 2$ 程度が適当だと思われる。

まず、LSA 全般の特性である高速性を検証した。500 回当たりの検索時間 (変換法 2,3 と同じ) を図 3 に示す。横軸は階数 k であり、20 までしか表示していないのは、データ行列 D のサイズが 20×40 なので、その階数は 20 以下であり、従って $k > 20$ は考えなくてよいからである。すべての $k (\leq 18)$ で、元の特徴ベクトルでの検索時間 0.92 秒よりも速くなっている。これは低次元化の効果である。

次に、本論文の主題であるロバスト化による検索精度の向上を検証した。ロバストでない場合 (変換法 1) の一致率を図 4 に、ロバストにした場合 (変換法 2) の一致率を図 5 に示す。変換法 2 での棄却半径の α は 2 とした。これは代表フレーム間距離を棄却半径とするということであり、一

般に妥当な設定といえる。 k を 6 や 8 辺りにすれば、ロバストでない場合 (図 4) でも、一致率は上記のインタセクションのときの 0.7 よりも大きく、図 5 は図 4 よりも更に一致率が向上している。なお、 k が 5 よりも小さいときに図 4 より図 5 のほうが一致率が低いことがあるのは、 k が小さすぎるとデータの分布を充分には表せないほど次元が下がり、正常な検索が不可能になるためと思われる。 $k = 2$ や 4 での逆転はこのような不安定な変動によるものと推測される。このことから、 k の値は 10 程度にするのが検索精度の面から安全であり、検索速度も比較的速い。

以上の実験では、ロバスト化によってクエリビデオと検索されたビデオとのグループの一致率が向上することを示したが、このような精度向上の汎用性について考察する。本実験では、特徴ベクトルとしてカラーヒストグラムを用いているので、同一グループのビデオの中に色の構成が大きく異なるものが含まれるような場合は、それらのビデオを正しく検索することは難しい。しかし、色の構成が異なる例外的なフレームの数があまり多くなければ、ロバスト化によってそれらのフレームを取り除いて検索することができる。すなわち、例外フレームが主要なクラスを構成しない程度に少なければ、それら例外フレームを除去できるので、提案手法によって検索の精度は向上すると考えられる。逆に例外フレームが全く含まれない場合は、提案手法はロバスト化なしの従来手法と同じ結果を出力するので、この場合も検索精度が低下することはない。本実験で用いたビデオデータに含まれる例外フレームとは、例えば、サッカーの映像で観客席の様子が短時間映し出される部分や、雪景色の映像に含まれる人のクローズアップのショットなどである。これらのショットを構成するフレームは、ビデオ全体の中では少数派であり、ロバスト化によって取り除かれる。

なお、ここで行った検索は類似検索であり、一致検索ではない。すなわちクエリが含まれるビデオの検索ではなく、クエリビデオとシーケンスが似ているビデオを検索するものである。各グループに含まれるビデオは、フレーム画像の重複がない互いに全く異なるシーケンスであり、本手法

はそのようなビデオデータについても有効性が実証され、したがって、一般的なビデオの類似検索においても、ここで提案したロバスト化が検索精度の向上に有効であると思われる。

最後に、画像の色空間と例外処理との関係について考察する。本実験では RGB 色空間を用いたが、予備実験では CIELAB 色空間での色の類似度を考慮した 2 次形式距離に基づく検索も行った。しかし、両者の間には大きな変化は見られなかったため、ここでは処理が簡単な RGB 色空間を採用した。実際、ある色空間を用いて距離が大きくなるフレーム対は、ほとんどの場合、他の色空間を用いてもフレーム間距離は大きいと考えられることから、色空間の選択によって例外処理の効果が左右されることはほとんどないと推測される。

5. む す び

ビデオの類似検索の 1 手法として、ベクトル量子化によりビデオを記号列で表して潜在意味解析で検索する手法を提案し、ベクトル量子化をロバスト化することにより検索性能が向上することを示した。本手法では、記号は単純なフレームクラスタであるが、記号表現を高度化して更に検索性能を高めるのが今後の課題である。

〔文 献〕

- 1) T. K. Landauer, P. W. Foltz and D. Laham: "Introduction to latent semantic analysis", Discourse Processes, **25**, pp.259-284 (1998)
- 2) D. Stan and I. K. Sethi: "Color patterns for pictorial content description", 17th ACM Symp. Applied Comput., pp.693-698 (Mar. 2002)
- 3) D. R. Heisterkamp: "Building a latent semantic index of an image database from patterns of relevance feedback", ICPR'02, pp.132-135 (2002)
- 4) 堀田政二, 宮原末治, 井上光平, 浦浜喜一: "潜在意味解析に基づくビデオの類似検索," 画像の認識・理解シンポジウム (MIRU 2002), **2**, pp.203-208 (2002)

いのうえ こうへい
井上 光平 1996 年, 九州芸術工科大学画像設計学科卒業。2000 年, 同大学大学院博士後期課程修了。現在, 九州大学助手。パターン認識, 画像処理, 画像検索, コンピュータビジョンに関する研究に従事。正会員。

うらはま きいち
浦浜 喜一 1980 年, 九州大学大学院工学研究科博士後期課程修了。同年, 同大学助手。1989 年, 九州工業大学助教授。現在, 九州大学教授。パターン認識, 画像情報処理の研究に従事。