SIFT 特徴量を用いた映像データに対する 人物検索システムの開発

丸西 立起† 獅々堀 正幹‡ 北 研二‡

†徳島大学大学院 先端技術科学教育部 システム創生工学専攻
‡徳島大学大学院ソシオテクノサイエンス研究部

〒770-8506 徳島県徳島市南常三島町 2-1

E-mail: {tatsuki, bori, kita}@is.tokushima-u.ac.jp

あらまし 大容量メディア, 動画共有サイトの普及により, 大量の映像データを視聴することが可能となった. しかし, 映像を視聴するユーザとしてはテキストのみに頼った検索方法では視聴したい映像, シーンを探すことが困難になった. そこで本稿では, 特定の人物が映るシーンを視聴したいといったニーズに応えるために, Scale-Invariant Feature Transform(SIFT)を用いた人物検索システムを開発した. また, TRECVID2010 の Instance search(INS)タスクの映像データを用いて, 従来の人物検索手法(固有顔を用いた人物検索システム)と, 提案システムとを比較した結果, 提案システムの有効性を確認した.

キーワード 情報検索, SIFT, 画像処理

Development of the instance retrieval system for the video data using the SIFT feature

Tatsuki MARUNISHI† Masamiki SHISHIBORI‡ and Kenji KITA‡

† ‡ The University of Tokushima

2-1 , Minamijousanjima-cho, Tokushima-shi, Tokushima, 770-8506 Japan

E-mail: {tatsuki, bori, kita}@is.tokushima-u.ac.jp

1. はじめに

近年, インターネット配信サービスの発展に伴い, ユーザは容易に大量の映像データを手に入れることが可能となった. しかし, 大量の映像データの中から必要なシーンを検索するのが困難なことから, 近年では映像検索技術に関する研究が盛んに行われている[1]. 様々な種類の映像検索技術がある中で我々が注目しているのは, 映像に映る人物の顔領域から得られる特徴を用いた検索手法である. この人物検索に関する研究は, TRECVID2010[2]でも Instance Search タスクとして実践されており, 近年, 注目をあびている分野である.

この人物検索技術, 特に人物の顔に着目した顔画像検索については, 従来, 固有顔(Eigen face)を用いて類似性を判定する手法[3]が主流であった. しかし, 固有顔は学習用に大量の顔画像を準備する必要があること, また, 姿勢や照明の変化に弱い点が問題となっている. その問題点を解決するために本稿では Scale-Invariant Feature Transform(SIFT)[4]を用いた人物検索システムを開発した. 本システムでは, まず, 映像データよりカットシーンを検出する. 次に, 検出したカットシーンに映る人物の顔領域であると思われる部分を抽出した後, 抽出した画像に対して SVM[5]を用いてノイズとなる顔画像(顔以外の部分が抽出された画像)を削除する. その後, ノイズが除去された顔画像に対して SIFT を用いて特徴量を取得し, データベースとする. 検索画像に対しても同様の処理を行い, データベース間で顔画像間の類似性を判定し, 結果として入力画像と同一の人物が映っているシーンを検索する.

以下, 2 章に SIFT について述べる. 3 章では人物検索システムについて説明し, 4 章で人物検索システムの有効性を検証するための評価実験とその結果を示す. 最後に 5 章においてまとめと今後の課題について述べる.

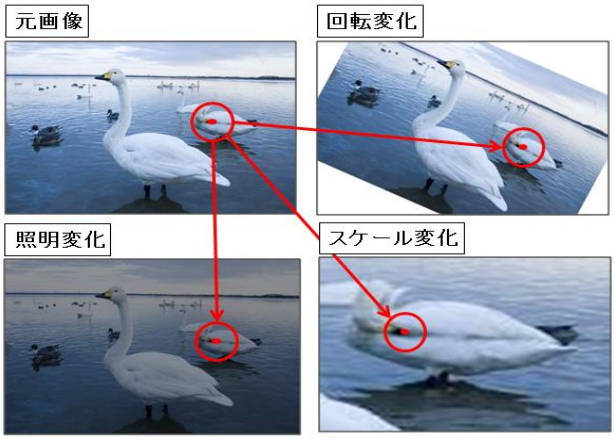


図 1 画像変化に対する SIFT 特徴点

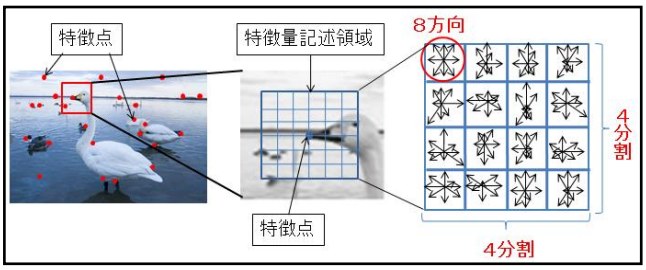


図 2 特徴量記述

2. Scale-Invariant Feature Transform(SIFT)

Scale-Invariant Feature Transform(SIFT)とは，画像より特徴点の検出と特徴量の記述を行うアルゴリズムである．SIFT 特徴点は，図 1 に示すようにそれぞれの画像変化に対して同じ場所に特徴点をとることができる．その特徴点周辺の勾配方向と勾配強度を用いることで，回転，スケール変化，照明変化等に頑健な特徴量を記述する．特徴量は，図 2 のように周辺領域を一辺 4 ブロックの 16 ブロックに分割し,ブロックごとに 8 方向の勾配方向ヒストグラムを作成する．よって，SIFT 特徴量は 4×4×8 = 128 次元の特徴量を作成する．

3. 人物検索システム

3.1. 人物検索システムの概要

映像データに対する人物検索システムの流れを図 3 に示す．まず，それぞれの映像データよりカットシーンと呼ばれるカメラの切り替わるシーンの検出を行う．次に，カットシーンに映っている人物の顔領域を OpenCV[6]を用い検出する．ただし，OpenCV により検出した画像には，顔以外の部分が検出されたノイズ顔画像が多く含まれている．そこで SVM を用いてノイズ顔画像をフィルタリングした後，それらの顔画像に対し SIFT 特徴量を取得し，それらを顔画像データベースとする．検索画像についても同様に，OpenCV を用

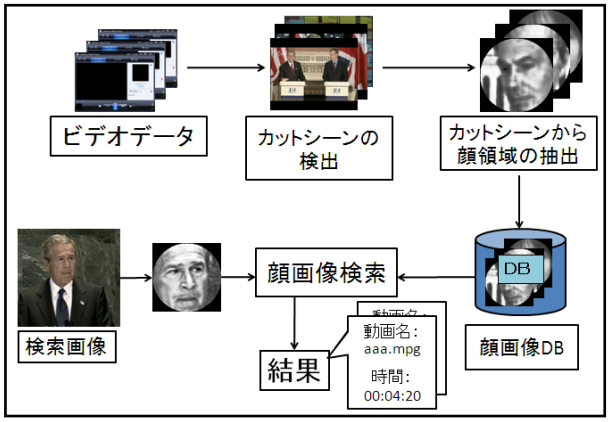


図 3 人物検索システムの流れ

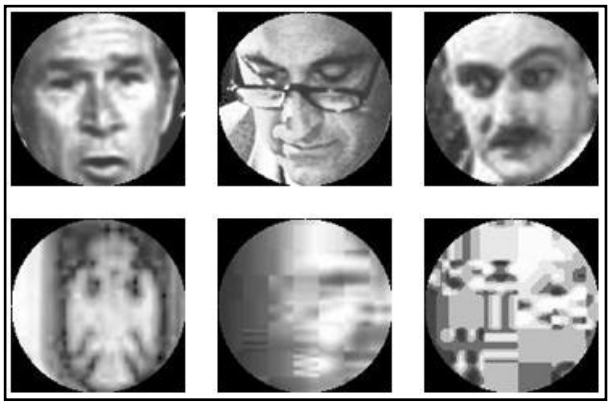


図 4 カットシーンから検出した顔画像例

いた顔領域の検出を行った後，SIFT 特徴量を取得する．そして，データベース間で検索を行い，結果を出力する．結果として，データベースに登録してある動画名と検索画像に映る人物が映っているシーンの時間を返す.以下に各処理の詳細を述べる．

3.2. 顔領域の検出

映像データから検出したカットシーンに対し，OpenCV を用い，Haar-Like 特徴量[7]により，顔領域の画像を検出する．検出した画像例を図 4 に示す.図 4 上段は正しく顔領域を検出することができた画像であるのに対し，下段の画像は顔以外の部分を検出したノイズ顔画像である．これらの画像は検索精度を低下させる恐れがあるため削除する必要がある．

3.3. ノイズ顔画像の除去

本システムでは，Haar-Like 特徴量を用い検出した顔領域の画像に対し，SIFT を使用した特徴量抽出を行っている．そこで，顔領域画像から SIFT 特徴点の検出された位置に着目した．図 5 に顔画像とノイズ顔画像からそれぞれ検出した SIFT 特徴点の位置分布を示す．図 5 から分かるように，顔画像に対する特徴点の位置

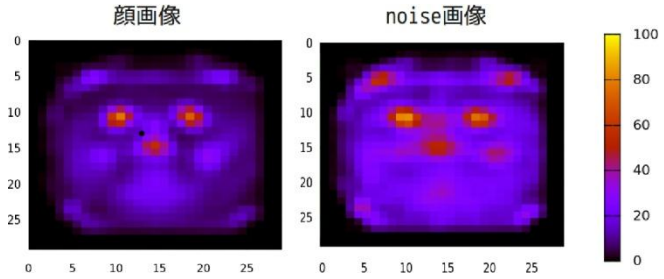


図 5 SIFT 特徴点分布

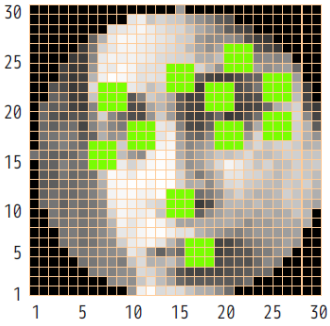


図 6 30×30 の SIFT 特徴点分布

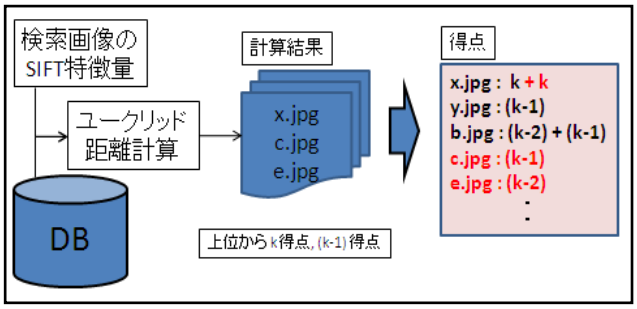


図 7 顔画像検索の流れ

分布は目,鼻の部分に集中しているが，ノイズ顔画像では顔画像よりも広範囲に SIFT 特徴点が分布している．これにより SVM を用いて SIFT 特徴点の位置から顔画像かノイズ顔画像であるかを学習し,削除している．具体的には，顔画像を図 6 に示すように 30×30 のブロックに分割し，各ブロック内に存在する SIFT 特徴点の頻度を各次元の値とした 900 次元の特徴量を用いた．また，顔画像から抽出される SIFT 特徴点には，ばらつきがあるため，頻度値ではなく 1(頻度あり)と 0(頻度なし)に 2 値化した特徴量を用いた．

3.4. 顔画像検索

図 7 に顔画像検索の流れを示す．まず，データベースと検索画像から抽出した顔画像の各 SIFT 特徴点に対して，128 次元の SIFT 特徴量の間でのユークリッド距離を計算する．それぞれの計算結果の上位より，k 得点，k-1 得点と得点を与える．最後にその得点を多く得

図 8 検索画像

たものから類似した画像とする．以下に例をあげる．例として，検索画像の顔画像に対して SIFT 特徴量を 3 つ取得したとする．それぞれの 128 次元の特徴量 1 つに対してデータベースの間との計算結果は 1 つずつ存在し，計 3 つの計算結果を得られることとなる．k = 10 としたとき，それぞれの計算結果上位の顔画像から 10 得点，9 得点と与え，0 得点になるまで続ける．得点を与えられた顔画像はその得点を保持し，保持した得点全ての総和で順位付けされる．図 7 の x.jpg は，k = 10 としたとき，20 得点を得たことになる．

4. 評　価

4.1 実 験 方 法

本手法の有効性を検証するために評価実験を行った．実験データには TRECVID2010 の InstanceSearch(INS)タスク[2]のデータを使用した．これらは，9 分から 60 分程度の動画 400 本からなり，カットシーンは約 41,000 枚，各カットシーンから顔画像約 17,000 枚を取得し，データベースとした．また，TRECVID2010 の INS タスクでは，人物検索以外に，特定のロゴマークやオブジェクトを検索するタスクも含まれていた．今回の実験では，人物検索タスクにのみ絞り，その中でも図 8 に示す 5 件の検索タスクに対して評価を行った．尚，今回用意したデータベースは正解画像数が検索画像に対して何枚含まれているのか分からないため，検索件数は上位 10 件，100 件，500 件，1,000 件までとし，その正解数を見た．また，従来法として固有顔による顔認識手法[3]と本手法を比較した．

4.2 実 験 結 果

図 9 から図 13 に従来法，提案手法の実験結果を示した.各図内の上部にあるブラウザ画像が検索結果となり，左上の画像から右に検索結果順の画像を表示しており，赤い丸で囲っているものが正解画像となる．また，各図内の下部にあるグラフが従来法と提案手法と

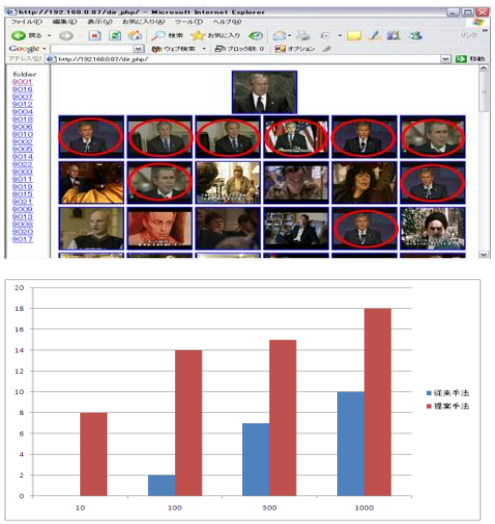


図 9 検索結果 1

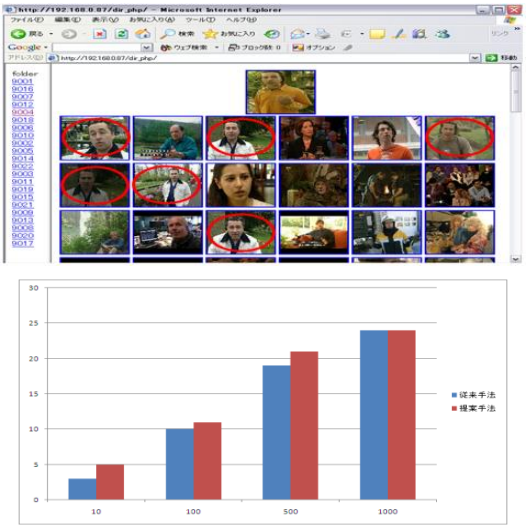


図 12 検索結果 4

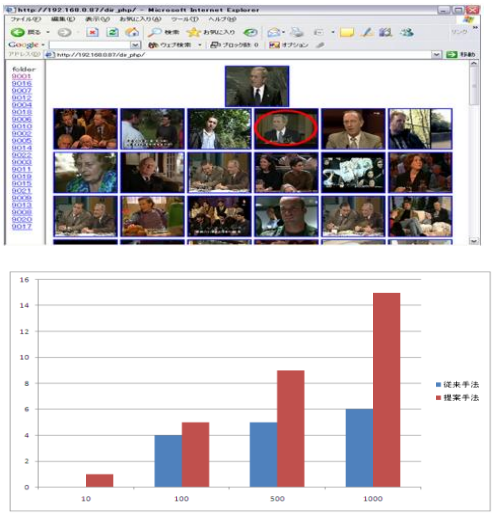


図 10 検索結果 2

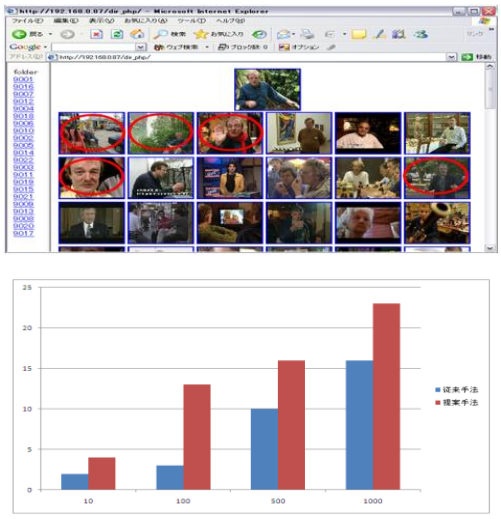


図 13 検索結果 5

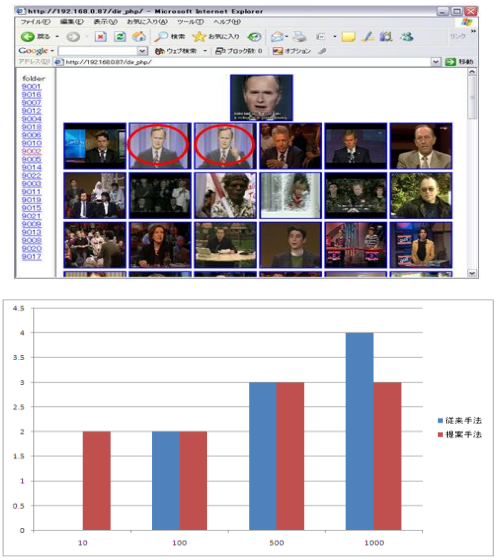


図 11 検索結果 3

の制度比較である．グラフの横軸は検索件数，縦軸は正解画像数である.赤いビンが提案手法，青いビンが従来手法となっている．ほとんどの検索画像について提案手法が従来手法よりも精度を向上させることができた.但し，図 11 の上位 1000 件以内の結果は従来手法のほうが多く正解画像を検索できていた．これは，提案手法では検索画像に対し，正解画像に含まれていた画像が，検索人物の若い時のものであったため，検索画像の顔にあるシワ，字幕に影響されたと考えられる．

図 9 と図 10 で使用している検索画像は同一人物であるが，図 9 のほうが上位に正解画像を多く検索することができている．これは，入力画像内の人物の姿勢(向き)に起因すると考えられる．図 9 の入力画像内の人物は正面を向いているため，検索結果には正面を向いている正解画像が多く検索できている．一方，図 10 の入

力画像内の人物は斜め方向を向いている．そのため，検索結果には，同じく斜め方向を向いている正解画像は検索されているが，図 9 で検索されていた正面を向いている正解画像は検索することができていない．このことから，今回の手法では人物の向いている方向に依存されてしまう傾向が強いと考えられる．

図 14 は，図 9 の検索結果 4 位(正解画像)と検索結果 7 位(不正解画像)に対して，検索画像との SIFT 特徴点の対応をとったものである．検索画像と検索結果 4 位の画像との対応点数 7 に対して，検索結果 7 位の画像との対応点数は 3 である．このことから，対応点を考慮した検索手法により精度向上が見込めるのではないかと考えられる．

検出された顔画像には顔領域のみ映っているわけではなく，背景部分が入り込んでいるものがある．背景部分の SIFT 特徴点は不要であると考えられるので削除することで精度向上が見込める．また，口周辺は変化が激しいと考えられるので，それらの部分についても SIFT 特徴点を削除することで精度向上するのではないかと考えられる．

図 15，図 16 に SIFT 特徴点を削除したときの結果を示す．図 15 より上位に正解件数が 1 件増えたことが分かる．このとき，新たな正解画像が増加しただけではなく，全体的に正解画像の順位が上位にきた．図 16 より上位 10 件での正解画像数は減少したが，上位 100 件以内での正解画像数では増加した．こちらも全体的に正解画像の順位は上がっており，この結果から口周辺部分を削除するのではなく，顔の変化が少ないと思われる部分，例えば目の周辺部分の SIFT 特徴量をよりみることによって精度向上するのではないかと考えられる．

また，本手法と従来法ともに正解画像を探せなかったという結果もある．これは，映像データからのカットシーン検出の時点で正解となる人物が映る画像を検出できなかったため，データベース内に正解画像が含まれなかったことが原因である．

5. まとめ

本稿では，SIFT を用いた映像データに映る人物を検索するシステムについて提案した．約 17,000 枚の人物の顔画像のデータベースを用いて行った評価実験では，従来法に対し提案手法のほうが正解画像数は多く，人物検索として検索精度の向上が見られた．また，SIFT 特徴点の対応点の考慮，不要な SIFT 特徴量の削除により精度向上が可能であると考えられる．また，データベース側に不備が確認されたため，今後は検索画像に対するデータベース内の正解画像数の確認を行う予定である．

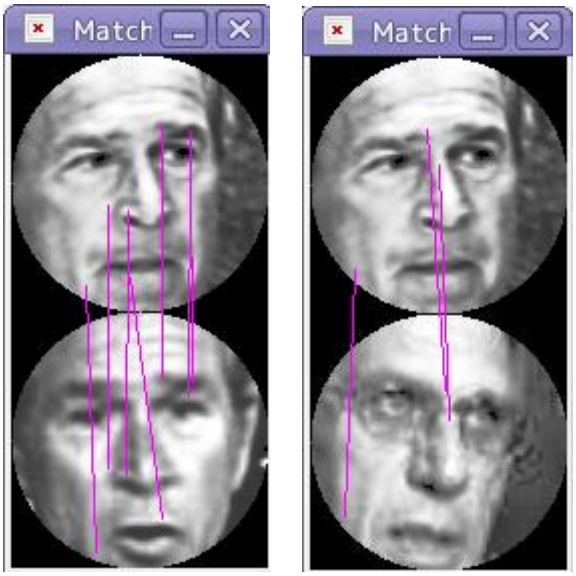
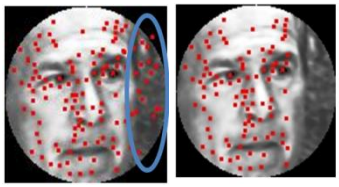
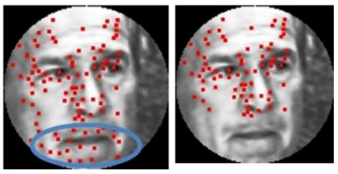


図 14　正解画像と不正解画像に対する検索画像との対応点



	削除前	削除後
上位10件正解数	7件	8件
上位100件正解数	12件	13件

図 15 背景部分の特徴点の削除結果



	削除前	削除後
上位10件正解数	8件	7件
上位100件正解数	13件	16件

図 16 口周辺部分の特徴点の削除結果

参　考　文　献

[1] Milan, P. and Willem, J. : Content-based video retrieval: A database perspective, Kluwer Academic Publishers(2003).

[2] http://www-nlpir.nist.gov/projects/tv2010/tv2010.html

[3] M. Turk and A. Pentland : Face Recognition Using Eigenfaces, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.586-591, 1991.

[4] D.G.Lowe, : Object recognition from local scale-invariant features, Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1150-1157, 1999.

[5] V.Vapnik, : The Nature of Statistical Learning Theory, Springer, (1995)

[6] http://opencv.willowgarage.com

[7] Rainer Lienhart and Jochen Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection”, IEEE ICIP 2002, Vol.1, pp. 900-903, Sep. 2002