卒 業 研 究 論 文

題　　目

局所特徴量を大規模に用いる Instance Search の有効性

知能メディア処理研究グループ

指 導 教 員　　黄瀬　浩一 教授

平 成　23　年（　　2011　年）度 卒 業

（No.　　1080107052　　）　　的﨑　伸彰

大 阪 府 立 大 学 工 学 部 知 能 情 報 工 学 科

局所特徴量を大規模に用いるInstance Searchの有効性

第3グループ　的﨑 伸彰

1. はじめに

TRECVID(Text REtrieval Conference VIDeo Retrieval Evaluation) とは，映像コンテンツの内容解析及び検索の高精度化を目的とする競争型国際プロジェクトである．TRECVID のタスクの一つに Instance Search(INS) がある．このタスクでは，ある特定の物体が写っている画像を用いて，その物体が映っている映像区間を探し出すことを目的としている．これは動画を対象とした特定物体認識に他ならない．

特定物体認識手法の一つに局所特徴量を用いたものがある．局所特徴量は，画像の局所領域から得られ，高い識別性を有している．しかし，1 枚の画像から数百から数千の特徴ベクトルが抽出されるため，大規模なデータを使用する INS ではメモリ使用量と処理時間が問題となる．そのため，従来手法では抽出した局所特徴量そのままではなく圧縮表現したものを用いることが多い．しかし，静止画における特定物体認識では，局所特徴量をそのまま用いて照合を行う方が良い結果が得られることが知られている．

そこで，我々は動画に対しても同じアプローチで試みる．このとき，メモリ使用量についてはメモリ容量の大きな計算機が利用可能なためそれほど大きな問題にならない．また，処理時間に関しては，ハッシュに基づく野口らの手法 [1] を用いることで対処する．この手法を動画検索に用いるにあたっては，高解像度画像を要求すること，色情報を用いないことが問題となる．これらの問題に対して，本手法では，画像拡大，色情報を含む局所特徴量の利用により対処する．実験の結果，画像拡大，色情報を含む局所特徴量の使用の両方で改善が見られた．

2. 手法

本手法では，映像から抽出した局所特徴量をデータベースに登録し，クエリ画像から抽出した局所特徴量用いて映像区間の検索を行う．まず，特徴抽出について説明する．INS で使用するデータは大規模のため，各動画から秒間 2 フレームを取り出しキーフレームとし，そこから局所特徴量を抽出する．その際，元の画像サイズでは物体領域から照合に有効な局所特徴量を抽出されないことがある．そこで，画像のサイズを拡大し，そこからも局所特徴量を抽出する．特徴量には色情報を含む OpponentSIFT 特徴量 [2] を用いる．OpponentSIFT 特徴量抽出では opponent color space の 3 つのチャンネルそれぞれから SIFT 記述子を用いて特徴を記述する．これによって各特徴点から 384 次元の色情報を持つ特徴量を得る．そして，抽出した特徴量を主成分分析によって次元削減する．

検索には野口らの手法を用いる．この手法は，ハッシュを用いた近似最近傍探索によって，局所特徴量を高速に対応づけるものである．動画の検索は，まず特徴ベクトルからハッシュ値を計算し，対応する ID と共に登録する．クエリから得られた特徴ベクトルに対してもハッシュ値を計算し，ハッシュ表にアクセスする．次に，そこから得られた特徴量に対して距離計算を行い，最も距離の小さい特徴ベクトルに対応する ID に投票する．この処理をクエリから抽出したすべての特徴ベクトルに対して行い，得票数の多い順に結果を出力する．

表 1: 認識結果 [%]

		DataBase			
		SIFT		OpponentSIFT	
		1 倍	1〜2 倍	1 倍	1〜2 倍
query	1 倍	5.9	5.2	8.9	9.0
	1〜2 倍	6.0	5.1	9.2	9.8
	1〜3 倍	6.3	5.2	9.5	10.1

3. 実験と考察

まず，本実験で用いたデータセットについて説明する．データベースとクエリには TRECVID2010 の INS のデータセットを用いる．映像は約 6 万件の短い動画で構成されており，全部で約 180 時間分ある．データベース作成の際，キーフレームのサイズを 1，2 倍にしたものから特徴量を抽出する．そして，1 倍のみ，1 倍と 2 倍の双方から抽出した特徴量を用いて 2 種類のデータベースを作成する．クエリには 9 つの物体を用いた．各物体には 3〜5 枚の画像が与えられており，これらを用いて物体に結果を求める．クエリも同様に，クエリ画像のサイズを 1，2，3 倍したものから特徴量を抽出する．そして，1 倍のみ，1 倍と 2 倍，1 倍から 3 倍から抽出した特徴量を用いて探索を行う．実験結果はクエリ毎に上位 1000 件の動画を求め，評価には MAP(Mean Average Precision) を用いた．また，OpponentSIFT 特徴量との比較として，色情報を含まない SIFT 特徴量を用いた場合についても実験を行った．

実験結果を表 1 に示す．TRECVID2010 の INS において最も良い結果は 3.3%であった．まず，OpponentSIFT 特徴量を使用した場合で，画像の拡大を行うと，キーフレーム拡大，クエリ画像拡大の両方で MAP の値に改善が見られた．これは，拡大した画像から特徴抽出することで照合に有効な特徴量を抽出できたためだと考えられる．次に，SIFT 特徴量と OpponentSIFT 特徴量を比較すると，SIFT 特徴量を用いた場合は 5.9%であったのに対して，OpponentSIFT 特徴量を用いることで 8.9%に改善された．クエリ毎に見ると，複数の色で描かれているクエリで大きな改善が見られた．全体として，SIFT 特徴量を使用した場合と画像拡大と OpponentSIFT 特徴量を使用した場合では最大約 4%の改善が見られた．

今後の課題として，本研究で用いた手法では検出できない物体や，特徴抽出そのものができない物体への対策が挙げられる．

参考文献

[1] 野口和人, 黄瀬浩一, 岩村雅一：“近似最近傍探索の多段階化による高速特定物認識”, 電子情報通信学会論文誌 D, **J92-D**, 12, pp. 2238–2248 (2009).

[2] K. E. A. van de Sande, et al.: “Color descriptors for object category recognition”, Proc. of CGIV2008, pp. 378–381 (2008).

目 次

図 目 次

表 目 次

第1章　はじめに

TRECVID(Text REtrieval Conference VIDeo Retrieval Evaluation) [3] とは，放送映像をはじめとした映像コンテンツを対象とし，その映像コンテンツの内容解析及び検索の高精度化を目的とする競争型国際プロジェクトである．米国標準技術局 (National Institute of Standards and Technology:NIST) の主催で開催されており，世界中から研究グループの参加を募り，参加者間で数百時間もの大規模映像アーカイブを共有すると同時に全員で同じタスクに挑戦し，その結果を比較評価することで研究水準の向上を目指している．2003年から開始され，今年まで毎年開催されている．TRECVID には 6 つのタスクが存在し，その中の一つに Instance Search(INS) がある．このタスクでは，ある特定の人物や物体が写っている画像を用いて，それらが映っている映像区間を探し出すことを目的としている．このとき，人物と物体では異なる手法を用いて検索をすることが多い．TRECVID2010 の結果では, 人物認識と比較して，物体認識が困難であるという結果が出ている [3] ．そこで，本研究では INS の中でも物体に限って評価を行う．検索対象を物体に限定するため，本研究で行う INS は動画を対象とした特定物体認識である．

特定物体認識手法の一つに局所特徴量を用いたものがある．この手法では，検索質問画像から抽出した特徴ベクトルとあらかじめデータベースに登録してある特徴ベクトルを照合し，対応する物体に投票する．次に，抽出した特徴ベクトル同士を照合し，特徴ベクトルに対応する物体に投票をする．最終的に，得票数の最も多い物体を認識結果とする．画像中の局所領域から特徴ベクトルを抽出するため，画像中の認識物体の位置や背景が変化しても高精度で認識可能である．しかし，1 枚の画像で数百から数千の特徴ベクトルが抽出されるため，大規模なデータを使用する INS では記憶に必要となるメモリ容量と局所特徴量の照合にかかる処理時間が問題となる．そのため，従来手法には画像全体から抽出する大域特徴量を使用したり，抽出した局所特徴量をそのまま使用するのではなく，圧縮表現することで特徴ベクトルの数を削減するものが多い．しかし，大域特徴量では画像中の物体の一部が隠れたり，背景が変化すると同じ特徴量を得ることができないため正し

く照合することができない．また，局所特徴量の圧縮表現についても，静止画における特定物体認識では，局所特徴量をそのまま用いて照合を行う方が良い結果が得られることが知られている．

そこで，我々は動画に対しても可能な限り局所特徴量の情報を残すアプローチで特定物体認識を行うことで高精度な検索を目指す．このとき，メモリ使用量については日々，技術の進歩によりメモリ容量の大きな計算機が安価で手に入るようになってきているため問題ではない．また，処理時間に関しては，野口らの手法 [1] を用いることで対処する．野口らの手法はハッシュやベクトルの各次元に対してスカラー量子化を用いることで，高速な処理とメモリ削減を実現するものである．

この手法を動画認識に用いるにあたっては，INS で使用する動画は低解像度であるため，照合に有効な特徴量を抽出できないという問題がある．そこで，フレーム画像を拡大し，そこからも特徴抽出することで照合に有効な特徴量を得る．また，色情報を用いないことも問題である．特定物体認識によく使用される局所特徴量として SIFT [4] や SURF [5] などが挙げられる．しかし，これらはグレースケール画像から特徴抽出する手法であるため，色情報を含んでいない．そこで，色情報を含む特徴量として動画中の物体検索に有効とされている OpponentSIFT 特徴量 [6] [7] を用いる．実験の結果，画像拡大と OpponentSIFT 特徴量を用いることで，約 10% の MAP(Mean Average Precision) を得た．また，TRECVID2010 において最もよかった結果と比較しても約 7% の改善が見られた．

第2章　INS(Instance Search)

本章では，TRECVID2010 における INS タスクの詳細について説明する．このタスクでは，ある特定の物体が写っている画像を用いて，その物体が映っている映像区間を探し出すことを目的としている．

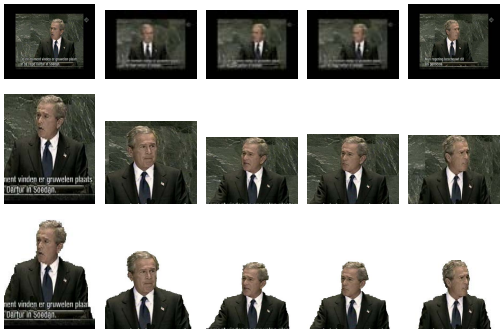
2.1　データベース

データベースには，Sound and Vision 2009 と呼ばれる映像が使用される．Sound and Vision 2009 はドラマやニュース，ドキュメントなど様々な番組が含まれる約 180 時間のオランダの放送映像で，著名な政治家，アナウンサー，劇中登場人物，ロゴ，特定の事物などがある程度の回数出現している．

Sound and Vision 2009 は映像の切り替わる部分で分割されている．分割された各々の映像を shot と呼ぶ．shot は全部で約 6 万件あり，それらに shot ID が付与されている．INS では目標物がどの shot に映っているかを調べる．

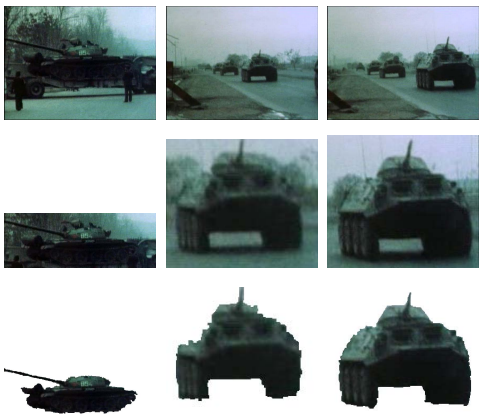
2.2　クエリ

各物体を検索する際の情報としてクエリトピックがある．クエリトピックの例を図 2.1 に示す．クエリトピックごとに 3〜5 枚の画像が用意されている．そして，その画像中の探したい物体領域の情報と，その画像の種別が「PERSON」(人物その人),「CHARACTER」(劇中の登場人物など．服装が不変などの特徴あり),「LOCATION」(場所),「OBJECT」(物体，ロゴなど) の文字列で与えられている．



(a) PERSON(George W. Bush)

(b) OBJECT(IKEA ロゴ)



(c) OBJECT(戦車)

(d) LOCATION(オランダの議会)

図 2.1: クエリトピックの一例

(それぞれ上から順に問い合わせの元フレーム画像，目標物の拡大画像，物体領域画像を表している.)

2.3 評価法

結果は，それぞれのクエリトピックに対して，クエリトピックが含まれると推定した shot ID の上位 1000 件を提出する．各クエリトピックの結果の評価には Average Precision(AP) を用いる．Average Precision の計算方法は以下の式以下のようになる．

$$A = \frac{1}{D_q} \sum_{1 \leq k \leq N} P_k \tag{2.1}$$

ここで，D_q は全正解数，N は正解が最後に現れた順位，P_k は k 番目の正解が検出された時点での適合率である．そして，すべてのクエリトピックの結果の

評価には，各クエリトピックの AP の算術平均である Mean AP(MAP) が用いられる．

第3章　従来手法

本章では，TRECVID2010 において最も良い成績を残した NII チームの手法 [8] について説明する．手法の全体的な流れを図 3.1 に示す．NII チームの手法では顔系と物体系で異なる特徴量を使用して照合を行う．特徴量には，局所特徴量をそのまま使用するのではなく，大域特徴や BoF(Bag of Features) を使用している．そのため，1 枚のフレーム画像から得られる特徴ベクトルは，各特徴抽出手法につき 1 個である．この手法で約 3.3% の MAP が得られる．本研究では物体認識に限るため，物体認識手法のみ詳しく説明する．

3.1　特徴抽出

3.1.1　前処理

まず前処理として，データベースの動画から 1shot あたり 50 枚の画像を切り出し，これをキーフレームとする．そして，得られたすべてのキーフレーム

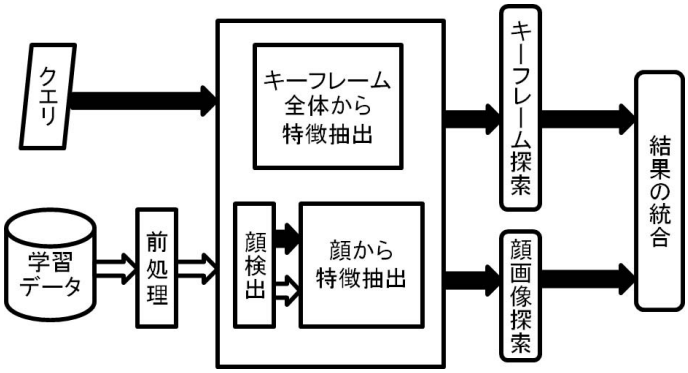


図 3.1: NII の手法の流れ

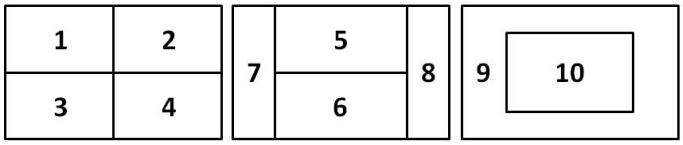


図 3.2: BoF の領域の分割

から大域特徴量の抽出と顔検出を行う．顔が検出された場合，顔領域を拡大した画像を切り出して，顔からも別の特徴量抽出を行う．

3.1.2 キーフレーム全体からの特徴抽出

物体の検索にはキーフレーム全体からの特徴量を使用する．キーフレーム全体からは 3 種類の特徴量を抽出する．1 つ目は色モーメントを用いる方法である．まず，キーフレームを 5×5 の領域に分割する．そして，それぞれの領域において HSV モデルを用いて，各パラメータの平均，分散，歪度を求める．すべての領域でもとめた値を連結することで $5 \times 5 \times 3 \times 3 = 225$ 次元の特徴量を得る．

2 つ目は LBP(Local binary patterns) [9] を用いた手法である．キーフレームを 5×5 の領域に分割し，各領域から 30 次元の値を得ることで $5 \times 5 \times 30 = 750$ 次元の特徴量を得る．

3 つ目は SIFT を用いた BoF 特徴量 [10] を抽出する．BoF では，複数の画像から抽出された特徴ベクトル群をクラスタリングする．そして，新たに画像から抽出された特徴ベクトル群をどのクラスタに属するかを比較し，その出現頻度のヒストグラムを特徴量として得る．まず，キーフレームから SIFT 特徴量を抽出する．次にキーフレームを図 3.2 のように 10 個の領域に分割する．そして，各領域で得られた SIFT 特徴量を基にして 738 次元の BoF 特徴を求める．最後に，すべての領域で求めた BoF を結合して，$10 \times 738 = 7380$ 次元の特徴量を得る．

3.2　検出方法と結果の統合

クエリ画像からもデータベースと同様に特徴量を抽出し，顔から求めた特徴とキーフレーム全体から求めた特徴でそれぞれ類似値を求める．最後に，顔から求めた検索結果は重みを 300, フレーム全体から求めた検索結果は重みを 1 で重み付けをして最終的な統合結果とする．

第4章　提案手法

本研究で用いた手法について説明する．まず特徴抽出について述べたあと，検索に用いた手法である野口らの手法について説明する．

4.1　特徴抽出

本節では，実験で使用するデータベースの作成方法と OpponentSIFT 特徴量 [7] について説明する．データベース作成の際は，動画を使用することと低解像度であることを考慮しながら作成する．そして，特徴量には色情報を含み，動画中の物体認識に有効とされている OpponentSIFT 特徴量を使用する．

4.1.1　キーフレーム抽出

INS では動画を使用する．動画は大量のフレーム画像で構成されており，このフレーム画像から特徴抽出を行う．しかし，すべてのフレーム画像から特徴抽出を行うのは，データ量，処理時間の観点からみても現実的ではない．また，時系列的に隣り合うフレーム画像にはほとんど差異は存在しないため，そこから抽出される特徴量も似通ったものになる．そこで，すべてのフレーム画像から特徴抽出を行うのではなく，毎秒２フレームを取り出し，それをキーフレームとする．そして，このキーフレームから特徴抽出を行う．

(a) 元画像　(b) チャンネル O_1

(c) チャンネル O_2　(d) チャンネル O_3

図 4.1: opponent color space への変換 [2]

4.1.2 画像の拡大

本研究で使用する手法では高解像度の画像を必要とする．しかし，INS で用いるデータセットは低解像度である．そのため，画像から特徴抽出をそのまま行っても，照合に有効な特徴量が得られないことがある．そこで，拡大画像からも特徴抽出を行うことで照合に有効な特徴量を得る．一般的に，画像の各辺を 2 倍にすると，得られる特徴量の数は約 4 倍となる．

4.2 OpponentSIFT 特徴量

画像認識の分野でよく使用される局所特徴量として SIFT [4] や SURF [5] といったものがある．これらはグレースケール画像から特徴抽出を行う手法のため，色情報を含んでいない．そこで，本研究では，色情報を含む局所特徴量で

動画中の物体認識に有効とされている OpponentSIFT 特徴量 [6] [7] を使用する．特徴点検出には Harris Laplace detector [11] を用いる．輝度値の変化に基づく特徴点の検出方法として，Harris detector がある．これに LoG(Laplacian of Gaussian) を用いたスケール探索の取り込むことで，スケール変化に耐性を持たしている．次に，特徴量の記述について説明する．まず，特徴抽出する画像を RGB 空間から

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{4.1}$$

によって Opponent color space [12] に変換する．

変換前と返還後の画像例を図 4.1 に示す．ここで得られたチャンネル O_1 と O_2 はそれぞれ赤と緑，黄と青の反対の色の組の情報を保持している．またチャンネル O_3 は HSV 色空間の明度に等しい．そして，Harris Laplace detector によって検出された特徴点より O_1 から O_3 のチャンネルごとに SIFT 記述子を用いて特徴抽出をする．SIFT 記述子とは回転や照明変化に頑健な記述子で高い識別性能を有している．最終的に，128 次元× 3 チャンネル= 384 次元の色情報を持つ特徴量を得る．本研究で使用する際は，384 次元そのまま用いると，メモリ使用量が膨大になることが問題になる．そこで，主成分分析を用いて次元削減を行う．

4.3 野口らの手法

本節では，本研究で検索に用いた手法である野口らの手法 [1] について説明する．この手法は，ハッシュやベクトルの各次元に対してスカラー量子化を用いることで，高速な処理とメモリ削減を実現するものである．はじめにデータ登録について述べたあと，検索方法について述べる．

4.3.1 データ登録

まず，参照動画のフレーム画像から d' 次元の特徴ベクトル $\boldsymbol{x} = (x_1, x_2, \cdots, x_{d'})$ を抽出する．この特徴ベクトル $\boldsymbol{x}$ の第 1 次元から第 d 次元 $(d \leqq d')$ までに対して，

$$u_j = \begin{cases} 1 \text{ if } x_j - \theta_j \geqq 0, (0 \leq j \leq d) \\ 0 \text{ otherwise}, \end{cases} \tag{4.2}$$

を適用することで，各次元を 2 値化したビットベクトル $\boldsymbol{u} = (u_1, u_2, \cdots, u_d)$ に変換する．ここで θ_j は，ハッシュ表に登録する特徴ベクトルすべての x_j の中央値である．そして，

$$H_{\text{index}} = \left(\sum_{i=0}^{d-1} u_i 2^i \right) \bmod H_{\text{size}} \tag{4.3}$$

によってハッシュ値を求め，shot ID とスカラー量子化によってデータ量を削減した特徴ベクトルを組にしてハッシュ表に登録する．ここで，H_{size} はハッシュ表のサイズである．特徴ベクトルのスカラー量子化を行う際は，特徴ベクトルの各次元の値である x_j を 2bit で表現する．そのため，量子化された値は 0，1，2，3 で表される．また，特徴ベクトルをハッシュ表に登録する際，すでに同じハッシュ値を持つ特徴ベクトルが格納されている場合 (衝突) がある．その場合は，チェイン法によって追加登録を行う．チェイン法とは連結リストを用いて，データを連結することで，ハッシュの一つの位置に複数のデータを格納する方法である．

特徴ベクトル登録の際に多数の衝突が起きる場合，距離計算の回数が増え，処理時間が増加してしまう．また，登録時に多数の衝突が起きる特徴ベクトルは，同じハッシュ値であり，類似している可能性が高い．これらの特徴ベクトルは識別性能が低いため，登録していても検索の際に誤対応の生じる原因となりうる． そこで，これらの問題に対処するため，リストとして登録される特徴ベクトルに対応する ID の種類に上限を設け，この上限を超える場合はリスト全体をハッシュ表から削除する．そして，そのハッシュ値への登録を行わないようにする．ここで，リストに登録する上限をリストとして登録した特徴ベクトルの個数ではなく，shot ID の種類の数としたのには理由がある．動画

は連続したフレームから構成されており，時系列的に隣り合うフレーム画像同士はほとんど差がない．そのため，それぞれのフレーム画像の同じ位置から抽出される特徴ベクトルも類似しており，同じハッシュ値に登録される場合が多い．仮に，リストの登録する上限を特徴ベクトルの個数にすると，同じような映像が続く shot では同じハッシュ値となる特徴ベクトルが多数出現し，削除されてしまう．しかし，shot 内に常に現れる特徴ベクトルが認識対象から抽出された特徴ベクトルでないとは限らない．ゆえに，特徴ベクトルに対応する shot ID の種類に上限を設けている．

4.3.2 検索

まず，クエリから得た各特徴ベクトル $\boldsymbol{q}$ に対して，データ登録と同様にハッシュ値を計算する．そして，ハッシュ表から特徴ベクトルを検索し，この特徴ベクトルの集合を $\boldsymbol{X}$ とする．次に，$\boldsymbol{q}$ をスカラー量子化したベクトルと，$\boldsymbol{X}$ に含まれるベクトルとのユークリッド距離を計算し，最近傍となる特徴ベクトル $\boldsymbol{x}_*$ を求める．そして，$\boldsymbol{x}_*$ に対応する shot ID に投票する．最近傍となる特徴ベクトルが複数ある場合には，それらすべてに対して投票処理を施す．ここで，shot 毎に抽出される特徴ベクトルの数 C_s は異なるため，特徴ベクトルの数が多い shot には投票される確率が高くなる．そこで，投票する際に $1/\sqrt{C_s}$ の重みを付けて投票を行う．クエリから抽出されたすべての特徴ベクトルに対してこの処理を行い，得票数の多い順に回答を出力する．

同じ物体を映していても，撮影条件によって特徴ベクトルの各次元の値は変化する．すると，ビットベクトルに変換する際，データベースに登録されている特徴ベクトルとは異なるビットベクトルになる可能性がある．ビットベクトルからハッシュ値を求めて検索を行うため，ビットベクトルが異なると正解を検索することができない．そこで，クエリから得られた特徴ベクトル q を用いて検索を行う際，この変化を考慮して検索を行う．まず，値の変動幅を e，クエリから得られた特徴ベクトルを $\boldsymbol{q} = (q_1, ..., q_d)$ とすると

$$|q_j - \mu_j| \leq e \tag{4.4}$$

を満たす場合は，q_j が閾値を超えている可能性がある．そこで，u_j だけでな

く，$u_j^{'} = 1 - u_j$ も用いてハッシュ値を計算し，ハッシュ表から特徴ベクトルを検索する．こうして得られた特徴ベクトルを $\boldsymbol{X}$ に加え，距離計算の対象とする．これを b 個の次元に対して行うことで，2^b 個のハッシュ値を計算することになる．

第5章　予備実験

本章では，OpponentSIFT 特徴量を何次元に削減するのが良いかを予備実験で確かめる．まず，実験条件について説明したのち，実験結果について述べる．

5.1　実験条件

実験条件について説明する．データベースとクエリには TRECVID2010 の INS のデータを使用した．ただし，データベースについてはすべてのデータを使用するのではなく，約 280 分，約 1800 shot の映像を用いた．クエリには 9 つの物体を用いた．クエリに使用した物体の画像例を図 5.1 に示す．ここで用いたクエリは「OBJECT」か「LOCATION」に分類されている．各物体には 3〜5 枚の画像が与えられており，これらを用いて物体毎の結果を求める．

次元数は，累積寄与率が 60% (22 次元)，70% (36 次元)，80% (60 次元)，90% (114 次元) となる次元に削減して実験を行った．実験結果は投票数の多い順に上位 1000 件を出力し，評価には MAP(Mean Average Precision) を用いる．

図 5.1: クエリの例

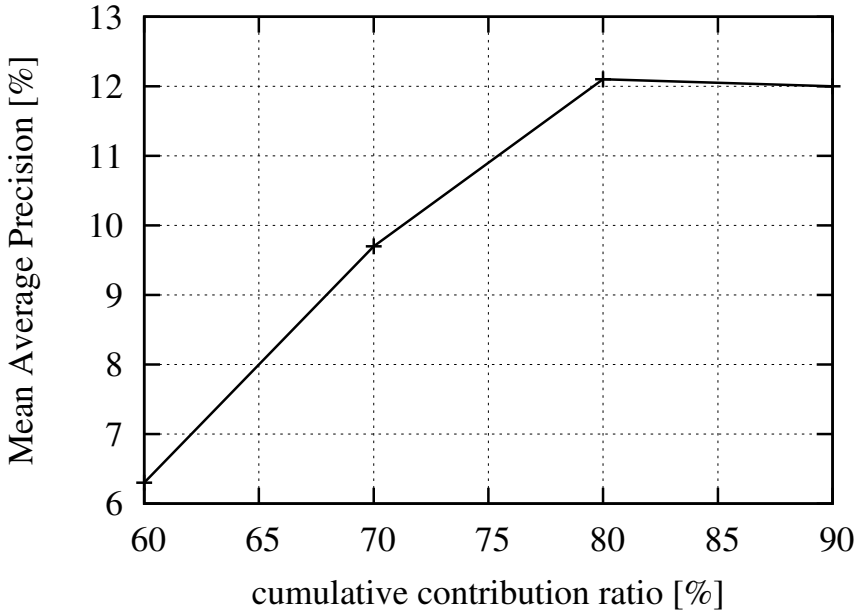


図 5.2: 累積寄与率と MAP の関係

5.2 実験結果

実験結果を図 5.2 に示す．実験結果より，累積寄与率が 80% (60 次元) のところで MAP が頭打ちになっていることがわかる．そのため，OpponentSIFT 特徴量を使用する際，主成分分析を用いて 60 次元まで次元削減しても十分な認識率が得られることが分かった．

第6章　実験

本実験では，画像拡大と色情報を含む特徴量である OpponentSIFT 特徴量を使用するとの有効性について検証した．まず，実験条件について説明したのち，実験結果と考察について述べる．

6.1　実験条件

本節では実験条件について説明する．データベースとクエリには TRECVID2010 の INS のデータを使用した．データベースの映像は約 6 万 shot で構成されており，全部で約 180 時間分ある．特徴量として，OpponentSIFT 特徴量を使用する．その際，抽出した特徴量は予備実験の結果より，主成分分析を用いて 384 次元から 60 次元に次元削減する．

データベース作成の際，キーフレームのサイズを 1，2 倍にしたものから特徴量を抽出する．そして，1 倍のみ，1 倍と 2 倍の双方から抽出した特徴量を用いて 2 種類のデータベースを作成する．

クエリには 9 つの物体を用いた．クエリに使用した物体は予備実験と同様である．各物体には 3〜5 枚の画像が与えられており，これらを用いて物体に結果を求める．クエリも同様に，クエリ画像のサイズを 1，2，3 倍したものから特徴量を抽出する．そして，1 倍のみ，1 倍と 2 倍，1 倍から 3 倍から抽出した特徴量を用いて探索を行う．

実験結果は投票数の多い順に上位 1000 件を出力し，評価には MAP(Mean Average Precision) を用いる．また，OppomnentSIFT 特徴量との比較として SIFT 特徴量を使用した場合についても実験を行う．SIFT 特徴量とは照明変化や回転，スケール変化に頑健な局所特徴量で，画像認識の分野で広く使用されている．OppomnentSIFT 特徴量とは異なり，SIFT 特徴量は特徴抽出の際

表 6.1: 認識結果 [%]

		DataBase			
		SIFT		OpponentSIFT	
		1 倍	1〜2 倍	1 倍	1〜2 倍
query	1 倍	5.9	5.2	8.9	9.0
	1〜2 倍	6.0	5.1	9.2	9.8
	1〜3 倍	6.3	5.2	9.5	10.1

に画像をグレースケール画像に変換するため，色情報を含んでいない．SIFT 特徴量も OpponenntSIFT 特徴量と同様に主成分分析を用いて，128 次元から 36 次元に次元削減する．この 36 次元とは，SIFT 特徴量において累積寄与率が約 80% となる次元である．

6.2 結果と考察

実験結果を表 6.1 に示す．TRECVID2010 の INS において NII チームの従来手法では 3.3% の MAP であった [8]．まず，OpponentSIFT 特徴量を使用した場合で，画像の拡大を行うと，キーフレーム画像拡大，クエリ画像拡大の両方で MAP の値に改善が見られた．キーフレーム画像の拡大については，IKEA や UMBRO のロゴ，オランダの議会で改善が見られた．これらは，他のクエリと比べて，色の境目がはっきりしている物体であるという特徴がある．そのため，キーフレーム画像を拡大した際にも，照合に有効な特徴量が抽出でき，スケール変化に対応できるようになったと考えられる．逆に横断歩道や戦車は画像がぼやけているため，キーフレームを拡大しても認識に有効な特徴量が抽出することができなくなったと考えられる．クエリ画像の拡大については，すべてのクエリで改善が見られた．これには 2 つの理由が挙げられる．一つは，キーフレーム画像拡大と同じく，クエリ画像を拡大することでスケール変化に対応できるようになったということである．もう一つは，本実験の評価法が

図 6.1: OpponentSIFT 特徴量によって改善したクエリ例

MAP であるということである．INS では正解 shot は複数存在する．MAP は上位に正解 shot が出現する方が良い結果となるため，正解 shot すべてに投票されるのが望ましい．しかし，投票数が少ないと，いくつかの正解 shot へ投票が偏ってしまったり，正解 shot 以外の shot の方が得票数が多くなる可能性がある．拡大したクエリ画像からも特徴抽出することで，得られる特徴量の数が増加する．これにより，各クエリトピックの投票数が増加する．その結果，正解 shot に投票される可能性が高くなり，結果が向上したと考えられる．ただし，クエリトピックの中には 3 倍まで拡大すると MAP が下がるものもあるため，何倍まで拡大すべきなのかはクエリトピックに依存するという問題点がある．

次に，SIFT 特徴量と OpponentSIFT 特徴量について比較する．SIFT 特徴量を用いた場合は 5.9% であったのに対して，OpponentSIFT 特徴量を用いることで 8.9% に向上している．クエリ毎に結果を見てみる．色特徴により改善が見られたクエリを図 6.1 に示す．これらのクエリは複数の色で塗り分けられており，色情報によって検出できるようになったと考えられる．逆に，色特徴により改善が見られなかったクエリを図 6.2 に示す．

図 6.2(a) のクエリはもともと白黒で描かれているため，OpponentSIFT 特徴量を使用しても色情報があまり得られなかったためと考えられる．また，図 6.2(b) のクエリはほとんど 2 色でクエリ画像が描かれている．OpponentSIFT 特徴量は色情報を含む特徴量である．しかし，画素の色の値を持っているわけでなく，各チャンネルでの濃淡をもとに特徴を記述している．そのため，使用

(a) クエリ例 1

(b) クエリ例 2

図 6.2: OpponentSIFT 特徴量によって改善しなかったクエリ例

図 6.3: 検索できないクエリ例

されている色の種類が少ないと効果が低いと考えられる．

全体として，SIFT 特徴量を使用した場合と画像拡大と OpponentSIFT 特徴量を使用した場合では最大約 4% の改善が見られた．また，TRECVID2010 における従来手法の結果と比べても，最大，約 7% の改善が見られた．しかし，すべてのクエリ画像が認識できているというわけではない．提案手法では検出できなかったクエリを図 6.3 に示す．これらのクエリは物体領域が画像全体に対して非常に小さく，色の境界がぼやけている．また，使用されている色も 2 色と少ない．そのため，画像を拡大しても対応できなかったと考えられる．

最後に，誤検出された shot の例を図 6.4 に示す．図 6.4(a) の裁判官の服では，同じ白黒の服装であるスーツが映っている shot を誤検出した．このよう

(a) 裁判官の服

(b) バス

(c) オランダの議会

図 6.4: 誤検出された shot 例

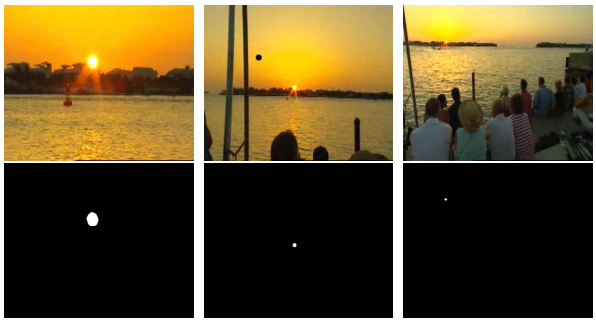
(左側にクエリ画像，右側に検出した shot を表している.)

に，見た目が酷似しており，色についても違いがないため，区別して検索するのは非常に難しい．そのため，より細かな形状の差異を表現できる特徴量を使用する必要がある．次に，図 6.4(b) のバスでは，バスの車体のペイントと同じものが服にもプリントされているため誤検出したと考えられる．このように，部分的に同じ部分が出てくると提案手法では誤検出してしまう．そこで，

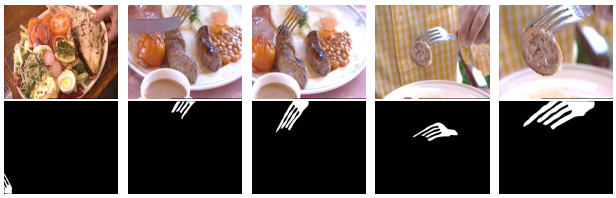
部分的な照合だけで結論を得るのではなく，他の部分も対応づくかどうかの検証が必要になる，また，図 6.4(c) のオランダの議会のように，検索対象が映っているにも関わらず，INS では間違いとなっているものもあった.

第7章　TRECVID2011での問題点

TRECVID2011では検索が困難であると予想されるクエリが出現している．TRECVID2011のクエリトピックの例を図3に示す．本研究で用いた特徴抽出手法では，図 7.1(a) 太陽のように単色で塗りつぶされており，模様が存在しない物体からは特徴を抽出することができない．そのため，物体領域を分割し各領域から特徴抽出をする，画像全体から大域特徴を抽出するなどの特徴抽出方法の改良も必要である．また，図 7.1(b) のフォークのように物体領域が一部分のみしか見えないものが存在するため，これらへの対策が必要である．



(a) 太陽



(b) フォーク

図 7.1: TRECVID2011のクエリトピック例

(上段に問い合わせの元フレーム画像，下段に目標物のマスク画像を表している.)

第8章　まとめと今後の課題

本論文では，従来手法のように大域特徴量を使用するのではなく，大規模な局所特徴量の照合という単純な手法により認識率の向上を目指した．そして，動画に対応するために画像拡大と色情報を含む特徴量である OpponentSIFT 特徴量を使用した．その結果，従来手法と比べると，最大 7% の改善が見られた．今後の課題として，本研究では検出することができなかったクエリや特徴抽出そのものが困難であろうクエリに対しての対策が挙げられる．

謝辞

本研究を進めるにあたって，直接御指導頂いた黄瀬浩一教授には，研究内容や論文の書き方，発表方法において多くの御指導，御助言を頂いたほか，プロジェクトへの参加等，活発な研究活動を導いて頂いたことを深く感謝致します．また，研究発表会等で様々な指摘および助言をしてくださった岩村雅一准教授，小島篤博准教授，内海ゆづ子助教に感謝致します．最後に，公私にわたり様々な支援及び助言をしてくださった知能メディア処理研究グループの皆様に感謝致します．

2012 年 3 月 9 日

参考文献

[1] Koichi Kise, Kazuto Noguchi, and Masakazu Iwamura. Robust and efficient recognition of low-quality images by cascaded recognizers with massive local features. *Proceedings of the 1st International Workshop on Emergent Issues in Large Amount of Visual Data (WS-LAVD2009)*, pp. 2125–2132, 2009.

[2] Margarita Bratkova, Solomon Boulos, and Peter Shirley. orgb: a practical opponent color space for computer graphics. *IEEE Computer Graphics and Applications*, Vol. 29, No. 1, pp. 42–55, 2009.

[3] 佐藤真一, 篠田浩一. 映像解析・検索評価ワークショップ TRECVID2010 の概要. 電子情報通信学会技術研究報告，PRMU2010-211, pp. 19–24, 2011.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, Vol. 60, pp. 91–110, November 2004.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, Vol. 110, pp. 346–359, June 2008.

[6] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1582–1596, 2010.

[7] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Color descriptors for object category recognition. In *European Conference on Color in Graphics, Imaging and Vision*, pp. 378–381, 2008.

[8] Le Duy-Dinh, Poullot Sebastien, and Satoh Shin'ichi. Baseline approach for instance search task: local region-based face matching and regional combination of local features (パターン認識・メディア理解). 電子情報通信学会技術研究報告, Vol. 110, No. 414, pp. 75–80, 2011-02-17.

[9] Timo Ojala, Matti Pietika"inen, and Topi Ma"enpa"a". Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, Vol. 24, No. 7, pp. 971–987, 2002.

[10] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Ce'dric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[11] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, Vol. 60, pp. 63–86, October 2004.

[12] J. van de Weijer and Th. Gevers. Boosting saliency in color image features. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pp. 365–372, Washington, DC, USA, 2005. IEEE Computer Society.