原著論文

Original Paper

ビデオ画像へのアノテーションのためのグループに基づいた画像検索手法

Group-Based Image Retrieval Method for Video Image Annotation

村林 昇
Murabayashi Noboru
筑波大学大学院ビジネス科学研究科
Graduate School of Business Sciences, University of Tsukuba, Tokyo
noboru@gssm.otsuka.tsukuba.ac.jp

倉橋 節也
Kurahashi Setsuya
（同　上）
kurahashi@gssm.otsuka.tsukuba.ac.jp

吉田 健一
Yoshida Kenichi
（同　上）
yoshida@gssm.otsuka.tsukuba.ac.jp

keywords: video, annotation, image retrieval, wavelet, SIFT

Summary

This paper proposes a group-based image retrieval method for video image annotation systems. Although the wide spread use of video camera recorders has increased the demand for an automated annotation system for personal videos, conventional image retrieval methods cannot achieve enough accuracy to be used as an annotation engine. Recording conditions, such as change of the brightness by weather condition, shadow by the surroundings, and etc, affect the qualities of images recorded by the personal video camera recorders. The degraded image of personal video makes the retrieval task difficult. Furthermore, it is difficult to discriminate similar images without any auxiliary information. To cope with these difficulties, this paper proposes a group-based image retrieval method. Its characteristics are 1) the use of image similarity based on the wavelet transformation based features and the scale invariant feature transform based features, and 2) the pre-grouping of related images and screening using group information. Experimental results show that the proposed method can improve image retrieval accuracy to 90% up from the conventional method of 40%.

1. は じ め に

パーソナルビデオカメラレコーダー (以下ビデオカメラ) が広く普及し，旅行や観光などで個人用ムービーを撮影することが多くなっている．一般にビデオカメラを撮影した日時情報は画像と一緒に記録再生できるが，撮影した観光地の風景画像が何であるかという情報は現状では記録再生できるものがない．手軽に撮影できる一方で，何を撮影したかというメタ情報はユーザが観光に行ったという記憶に基づいて後で手作業で作成するしかなく撮影情報が何もないビデオコンテンツが沢山たまってしまうことになる．

このような課題を解決する一つの手段としてビデオ画像に自動的にアノテーションを付与することを考えた．ビデオ画像にアノテーションを付与することにより，旅行先における思い出の画像を検索したり，ビデオコンテンツを整理することが容易になる．さらに，撮影したビデオの代表画像やタイトルを決めることなどができるようになり，個人的な旅行の画像アルバムを効率的に作成することも考えられる．このように，ビデオ画像へのアノテーションの付与は，コンテンツに新たな付加価値を与える様々なアプリケーションを創造する．

我々はビデオ画像にアノテーションを付与する手法として，あらかじめアノテーションと対応した画像が蓄積された画像データベースを用いる方法を考えた．ビデオ画像と類似した画像を画像データベース内から検索し，検索された画像に対応したアノテーションをビデオ画像に付与する方法である．画像データベースとしては WWW 上の観光や旅行案内情報を利用する．WWW ページに掲示された観光名所の代表的な画像をそのまま画像データベースの画像として用い，ページタイトルなど WWW ページ内の文字情報 (これらは通常その観光名所の説明文である) をアノテーションとして利用する．ビデオ撮影時にはユーザが検索処理に利用するための画像を撮影しておくこととする．具体的には各観光地の代表的な景色やランドマーク的な建物を旅の記念に撮影しておき，ヒ

ントとしてアノテーション処理時にそれらの画像を選択してシステムに提示することを想定する．

しかし，旅行先でのビデオ撮影を考えた場合，必ずしも良好な天候や安定した撮影環境のもとでビデオを撮影できるとは限らない．このような条件下で撮影されたビデオ画像は，天候の状況による明るさの変化や周囲の陰影などの影響を受け，画質が良好でない場合が多い．またビデオカメラやデジタルカメラの機器間の相違により，撮影画像における色の整合性が取れないこともある．例えば異なったビデオカメラやデジタルカメラで同じ風景を撮影した場合でも，撮影した画像を再生して見た場合に，画像間で明るさや色の特性が同じではないことが起こりえる．このようなことから，本研究では撮影条件などが校正されずに同一対象の画像を撮影していても結果として画質の違いにより画像特徴が異なっている画像の検索手法について検討を行った．

さらに，画像検索手法単独では類似した画像が検索対象の画像データ内に複数存在した場合にそれらを区別することは困難で，補助的手法を用いる研究が多かった [Wang 06] [長坂 98]．例えば，よく知られた観光地である浅草と奈良に行った場合を考えてみる．浅草の浅草寺における五重塔と奈良の法隆寺における五重塔をビデオで撮影した場合に，両方の撮影画像は類似しており，浅草または奈良のどちらの場所で撮影したかという情報が分からないとそれら２つの画像を区別することは難しい．

このような課題すなわち，1) 明るさなどの撮影条件が校正されないことによる画質の違いのために，同一対象の画像を撮影していても結果として画像特徴が異なっている画像の検索が困難なこと，2) 本質的に画像特徴だけでは類似した画像の区別をすることは困難なこと，の２つに対応するために，本論文ではビデオ画像へのアノテーションのためのグループに基づいた画像検索手法 (グループ法) を提案する．

提案するグループ法の特徴は，1) wavelet 変換に基づいた領域ベースの大局的画像特徴量 [Wang 97] と SIFT(Scale Invariant Feature Transform) 特徴 [Lowe 04] に基づいた局所的画像特徴量を組み合わせて利用することと，2) 事前に検索対象の画像を関係するものごとにグループ化しておき，まず対応した類似画像を多く含むグループを選択し，次に選択したグループ内から類似画像を選択すること，である．第１の特徴は，天候の変化など様々な撮影環境下で撮影された同一対象の画像の間で検索を行う場合における画質の違いに対応することを目的とし，第２の特徴は個々の画像特徴による検索処理とは別の補助的手法を用いた検索性能のさらなる改善を目的としている．

以下．第２章で関連研究について概観した後，第３章で提案する画像検索手法を説明する．第４章で実験結果について報告し，第５章で考察を行う．最後に第６章でまとめを行う．

2. 関 連 研 究

ビデオ画像や静止画像へのアノテーションに関しては従来から様々な手法の研究が行われている．例えば，Abowd らは，アノテーション処理をサポートするため drag-and-drop 操作を用いる手法を提案している [Abowd 03]．Song らの研究 [Song 05] は，機械学習の技術を適用した半自動のアノテーション処理を試みている．Yu らは，画像検索技術を用いた手法を提案している [Yu 06]．

我々は，Yu らの画像検索技術を用いたアノテーションの研究に着目しその改良を試みた．Yu らの研究の課題は，目標画像に対して正確なアノテーション処理を行うことが困難なこと，すなわち画像検索性能が良くないことである．その理由として，検索対象の画像のスケーリングや撮影方向が異なると画像検索性能を上げることが難しいこと，気象天候状況など画像の撮影条件が異なると検索対象の類似した画像でも対応する色などの画像特徴が異なること [Narasimhan 00] がある．

画像検索の性能向上に関係する研究としては [柳井 04, Rubner 99] などがある．柳井は画像を複数の固定した領域ブロックに分割した上で，その領域ブロックごとに色に基づいた画像特徴量を抽出し，EMD(Earth Mover's Distance)[Rubner 98] を用いて画像間の類似度を計算することで WWW ページから収集した画像を検索する手法を提案している [柳井 04]．Rubner らは特徴ベクトルを生成するのに Gabor フィルターを適用し，画像データとして Corel Stock Photo Library を用いた画像検索手法について研究を行っている [Rubner 99]．

これらの手法は画質の違いが少ない画像を対象としている．例えば前者の WWW ページにおける画像を考えた場合，画像は他の第３者に公開することを前提にしており，ユーザは画像を選別して WWW 上に公開している．すなわち [柳井 04] における WWW 上の画像は，一般ユーザが撮影した画像であるにしても，比較的画質が良好で画質の違いが少ない画像であると言える．また後者が実験に用いた Corel Stock Photo Library は商用のもので，画質は良好である．

しかし，我々の研究対象としている典型的な民生用ビデオカメラで撮影した画像は，天候状態や撮影の周囲環境の陰影，その他の撮影条件などの影響により，そのような WWW 上の画像や商用の画像と比較して同一対象の画像であっても画質に違いがある．本研究では，このような画質が違うことが検索性能に及ぼす影響を，予備検討の結果 [村林 07] が良かった wavelet 特徴と，近年着目を集めている SIFT 特徴 [Lowe 04] を組み合わせて使い対策することを考えた．

更に，浅草の浅草寺における五重塔と奈良の法隆寺における五重塔のような似た風景画像の識別・検索は本質的に困難な問題である．このような場合に画像検索性能を改善する方法として，画像処理以外の補助的な手法を

利用する研究もある．例えば Wang らは，テキストに基づいた検索技術によって画像検索性能を改良した研究を行っている [Wang 06]．[長坂 98] は，補助的手法として特徴の変化パターンを利用して映像シーンの検索性能を改良している．しかし，ビデオ画像のアノテーションを目的とした場合，どのような補助的な手法が有効であるかの検討はされていない．

以上述べて来たように，ビデオ画像のアノテーションエンジンとして従来手法を単純に用いることはできない．我々は補助的な手法として，関連性のある検索対象の画像をグループ化して，各最小画像間距離の総和を計算することで検索を行うグループに基づいた画像検索手法を検討し，予備検討結果を報告した [村林 07]．以下本稿では予備検討結果をベースに画像処理自体の改善方法と，より多くのデータを用いての実験結果を報告する．

3. ビデオ画像へのアノテーションのための画像検索手法

この章では，ビデオ画像にアノテーションを付与するための画像検索手法を提案する．この画像検索手法の目的は，WWW 上における撮影場所情報とそれに対応する画像を参照 WWW 画像として用いて，様々な観光名所で撮影されたビデオ画像と類似する参照 WWW 画像を見つけることで，使い方として以下を想定している．

(1) 参照 WWW 画像は該当する観光名所を含む検索キーワードに基づいて WWW 画像検索エンジンにより収集する．そのような WWW ページの画像は，国内外の自治体，政府観光局，観光業者など WWW ページの所有者によって選択された観光名所の代表的な画像と想定する．

後で説明する実験のために画像収集した経験によれば，一つの観光地で紹介されている観光名所はおむね 4〜5ヶ所程度であった．そこで以下の議論では一つの地域に含まれる観光名所は５つと想定し，各参照 WWW 画像グループごとに観光名所に対応する画像が５枚あるものとする．

(2) ユーザは Graphical User Interface を通して撮影したビデオコンテンツから複数のビデオ画像の選択を行う．ユーザが選択したビデオ画像は代表的なものであり，対応する参照 WWW 画像が存在しやすいものと想定する．

撮影されたビデオ画像には人物が撮影されたものや，観光名所以外の画像が含まれる場合もあるが，そのような画像を除外した選択をユーザがすることを想定する．すなわち，ユーザは観光名所の代表的な景色やランドマーク的な建物を旅の記念に撮影しておき，撮影したビデオ画像の中からそれらの画像を選択してヒントとしてアノテーション処理時にシステムに提示することを想定する．

(3) ユーザが選択したビデオ画像に類似した参照 WWW 画像を検索し，参照 WWW 画像を収集した際に使用した検索キーワードと WWW ページのテキスト情報をビデオ画像のアノテーションとして用いる．

また予備的な検討 [村林 07] では，従来の研究 [柳井 04] で用いられた色特徴に基づいた画像検索手法は，ビデオカメラで撮影された様々な天候条件下における同一対象の画像であっても画質が違う画像に対しては検索性能が良好ではなかった (5·1 節を参照)．我々の提案手法ではこのような問題に対応するために wavelet 変換 [Wang 97] に基づいた特徴 (以下 wavelet 特徴) と SIFT 特徴 [Lowe 04] の２種類の特性が異なる画像特徴量を組み合わせて用いている．

さらに，撮影された代表的な画像には五重塔のように複数の類似した観光名所の画像が含まれる可能性もあるが，観光地によって撮影された代表的な画像の組み合わせが異なるので，撮影された観光地ごとに画像のグループ化を行う処理によってこのような類似画像を排除することを考えた．

以下，提案手法について詳しく説明を行う．

3·1 wavelet 特徴による画像間距離

はじめに提案手法で用いる２種類の画像特徴の一つである wavelet 特徴による画像間距離について説明する．wavelet 特徴による画像間距離は図 1 に示す処理プロセスに従い，画像全体からその大局的な特徴に基づく距離を計算するものである．

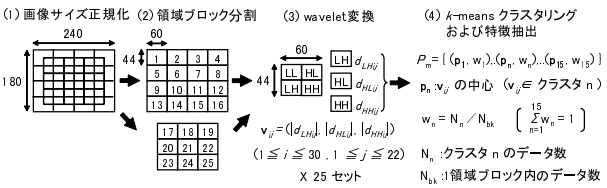


図 1 wavelet 特徴による処理フロー

(1) はじめに，参照 WWW 画像とビデオ画像をビットマップフォーマットに変換し 240 × 180 画素の画像サイズに正規化した上で (図 1 (1))，YCbCr(Y: 輝度信号，Cb,Cr: 色差信号) 色空間のデータに変換する．

(2) 次に，正規化した画像を 60 × 44 画素の 25 の領域ブロックに分ける (図 1 (2))．

(3) それぞれの領域ブロックにおいて，Y信号データを用いて，Daubechies-4 wavelet 変換処理 ([Wang 97]) を行い LH, HL, HH 成分から特徴ベクトル $\mathbf{v}_{ij}$ を生成する (図 1 (3))．具体的には，

$$\mathbf{v}_{ij} = (|d_{LHij}|, |d_{HLij}|, |d_{HHij}|) \qquad (1)$$
$$(1 \le i \le 30, \quad 1 \le j \le 22)$$

ここで，$d_{LHij}, d_{HLij}, d_{HHij}$ は，wavelet 変換にお

ける LH, HL, HH 成分で，i, j は，それぞれの wavelet 変換成分の次元である．

(4) 各領域ブロックにおいて，$\mathbf{v}_{ij}$ を用いた k-means クラスタリング処理を行い，領域ブロックごとの画像の特徴 P_m $(1 \leq m \leq 25)$ を，それぞれのクラスタの中心値と，クラスタサイズの割合を用いて以下の式で計算する (図 1(4))．クラスタ数は [柳井 04] にならって 15 とした．

$$P_m = \{(\mathbf{p}_1, w_1), (\mathbf{p}_2, w_2) ... (\mathbf{p}_{15}, w_{15})\} \quad (2)$$

ここで，$\mathbf{p}_n$ $(1 \leq n \leq 15)$ はクラスタ n の中心で

$$\mathbf{p}_n = (\frac{\sum |d_{LHij}|}{\substack{\text{クラスタ } n \text{ に属} \\ \text{する } \mathbf{v}_{ij} \text{ の数}}}, \frac{\sum |d_{HLij}|}{\substack{\text{クラスタ } n \text{ に属} \\ \text{する } \mathbf{v}_{ij} \text{ の数}}}, \frac{\sum |d_{HHij}|}{\substack{\text{クラスタ } n \text{ に属} \\ \text{する } \mathbf{v}_{ij} \text{ の数}}}) \quad (3)$$

上式の中で $\sum |d_{LHij}|, \sum |d_{HLij}|, \sum |d_{HHij}|$ は，いずれもクラスタ n に属する $|d_{LHij}|, |d_{HLij}|, |d_{HHij}|$ それぞれの総計を表す．

w_n は，クラスタサイズの割合で

$$w_n = \frac{\text{クラスタ } n \text{ に属する } \mathbf{v}_{ij} \text{ の数}}{\mathbf{v}_{ij} \text{ の総数}} \quad (4)$$

である．

(5) 次にビデオ画像 P における番号 m の領域ブロックと，参照 WWW 画像 Q における番号 m の領域ブロックとの EMD[Rubner 98] を EMD(P_m,Q_m) とし，P と Q 2 つの画像間の距離 EMD(P,Q) を同じ領域ブロック番号同士の EMD 距離の総和とした．

$$\text{EMD}(P,Q) = \sum_{m=1}^{25} \text{EMD}(P_m, Q_m) \quad (5)$$

(6) 最後に SIFT 特徴と組み合わせて用いるために，データ範囲を [0,1] とし正規化する．具体的には，ビデオ画像 P に対して最大の EMD を持つ参照 WWW 画像の EMD 値を dw_{max}，最小の EMD 値を dw_{min} とすると，ビデオ画像 P と参照 WWW 画像 Q との wavelet 特徴による正規化 EMD dw_{PQ} を以下のように定義する．

$$dw_{PQ} = \frac{\text{EMD}(P,Q) - dw_{min}}{dw_{max} - dw_{min}} \quad (6)$$

3·2 SIFT 特徴による画像間距離

画像特徴として，前述の wavelet 特徴の他に局所的特徴量である SIFT 特徴を用いた．SIFT 特徴による画像間距離は，図 2 に示すように，画像に含まれる特徴点 (keypoint) 周辺の局所的な特徴に基づく距離を計算するものである (詳細については [Lowe 04] を参照されたい)．なお，以下で説明する (1)〜(4) の処理は，ビデオ画像と参照 WWW 画像それぞれに対して同じ処理を行うため，簡単のために一方の画像について説明する．

(1) wavelet 特徴の場合と同じく画像を 240 × 180 画素の画像サイズに正規化する (図 2 (1))．

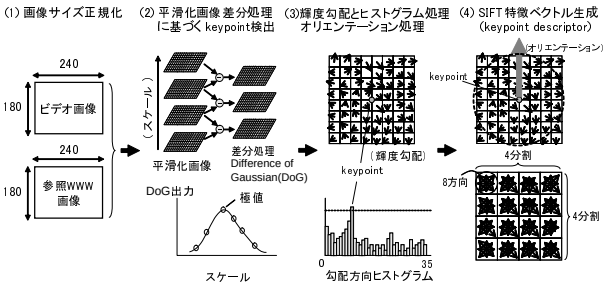


図 2 SIFT 特徴による処理フロー

(2) 正規化された画像に対して，平滑化パラメータ（スケール）を変えて差分処理 (Dog:Difference-of-Gaussian 処理) を行い，変化量の大きくなる点 (極値となる点) を検出し keypoint とする (図 2 (2))．

(3) 求めた keypoint の周辺領域から，輝度勾配と勾配方向を求めてヒストグラム処理を行い，ピークを各 keypoint におけるオリエンテーションとする (図 2 (3))．

(4) keypoint の周辺領域をオリエンテーションに従って回転させる．周辺領域を 4 × 4 の 16 の小領域に分け，各小領域で 8 方向の輝度勾配方向のヒストグラムを求める．ヒストグラムの各方向が特徴ベクトルとなり各特徴ベクトルの長さをベクトルの総和で正規化し 128 次元の SIFT 特徴ベクトル (keypoint descriptor) とする (図 2 (4))．

(5) ビデオ画像 P と参照 WWW 画像 Q の各画像との間で特徴ベクトルのマッチング処理を行う．マッチング処理では SIFT 特徴ベクトル間のユークリッド距離があらかじめ予備実験により定めたしきい値より近いものを対応する特徴ベクトルと判定する．

(6) wavelet 特徴による EMD の場合と整合性を持たせるためビデオ画像 P と参照 WWW 画像 Q との keypoint 対応数 n_{PQ} による画像間距離をデータ範囲 [0,1] で対応する特徴ベクトルの数 (マッチング数) が大きい場合には画像間距離が近くなるように定義する．具体的には，ビデオ画像 P に対して最大のマッチング数を持つ参照 WWW 画像のマッチング数を n_{max}，最小のマッチング数を持つ参照 WWW 画像のマッチング数を n_{min} とし，ビデオ画像 P と参照 WWW 画像 Q との SIFT 特徴による正規化画像間距離 ds_{PQ} を以下のように定義する．

$$ds_{PQ} = 1 - (\frac{n_{PQ} - n_{min}}{n_{max} - n_{min}}) \quad (7)$$

3·3 2 つの特徴による画像間距離データの重み付け加算

最終的には画像間距離は前述の 2 つの距離を重み付け加算して用いた．具体的には重み係数 k を考え，

$$dws_{PQ} = k \cdot ds_{PQ} + (1-k) \cdot dw_{PQ} \quad (8)$$

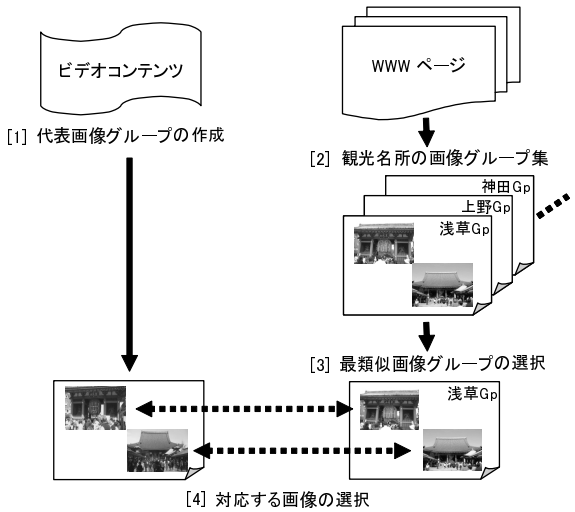


図 3 グループに基づいた画像検索処理

```
Pg=処理対象のビデオ画像グループに含まれる画像の集合
Qg=処理対象の参照 WWW 画像グループに含まれる画像の集合
n = min(Pg に含まれる画像数 np,Qg に含まれる画像数 nq)
for (i=0; i<n; i++) {
    Find 最小画像間距離を与える Pg に含まれる画像 p と
                        Qg に含まれる画像 q の組み合わせ
    tmpP[i] = p
    tmpQ[i] = q
    Pg = Pg から p を削除
    Qg = Qg から q を削除 }
for (a=0,i=0; i<n; i++)
    a += (tmpP[i] と tmpQ[i] 間の画像間距離)
for (b=0,i=0; i<n; i++)
   for (j=0; j<n; j++)
       b += (tmpP[i] と tmpQ[j] 間の画像間距離)
a/b を Pg と Qg 間の距離とする
```

図 4 画像グループ間の距離計算方法

とした．$k\,(0 \leq k \leq 1)$ はチューニングが必要なパラメータであるが，この値については 5 章で述べる．

以上の画像間距離の定義の中で，wavelet 特徴を用いた画像間距離 dw_{PQ} は代表的な類似研究である [柳井 04] で使われている色に基づいた画像特徴の代わりに wavelet 特徴を用いた予備検討 [村林 07] で提案したものである．また SIFT 特徴に基づく画像間距離 ds_{PQ} は，予備検討の結果をベースに本論文で新しく提案したものである．最終的な画像間距離 dws_{PQ} は 2 種類の特性が異なる画像特徴量に基づく画像間距離 dw_{PQ}, ds_{PQ} を組み合わせ検索性能を向上している．

3·4 グループに基づいた類似画像検索

図 3 にグループに基づいた画像検索処理 (グループ処理) のアイデアを示す．はじめにユーザは，撮影されたビデオ画像から代表的な画像 n_p 枚を選択し，ビデオ画像グループを作成する (図 3 [1])．次にあらかじめ幾つかの WWW ページから作成しておいた様々な観光地における観光名所の画像グループ (G_p) を画像グループ集とし (参照 WWW 画像グループ: 図 3 [2])，その中から作成したビデオ画像グループに最も類似した画像グループを選択する (図 3 [3])．最後に選択した画像グループの中から，画像グループの最小距離を与えた検索対象のビデオ画像と対応する参照 WWW 画像を検索する (図 3 [4])．

このグループ処理においてグループ間の距離は図 4 に示す手順で求めた．本研究では各画像グループは観光名所の代表的な風景画像からなることを想定している．また，代表的風景どうしは互いに似ていない画像を持つことも想定している．このような想定が実データで成立していれば，ビデオ画像と参照 WWW 画像の間の画像間距離を記録した行列は，図 5(3) に示したように対角成分が他の成分より小さな行列になるはずである．図 4 に示す手順は，そのような関係が成り立つ画像グループ間の

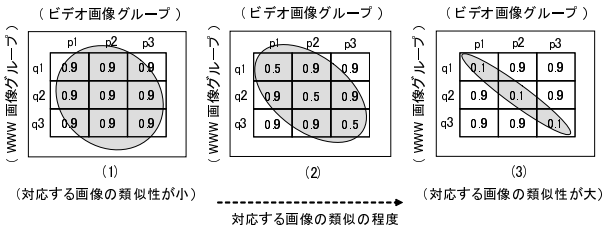


図 5 ビデオ画像と参照 WWW 画像の間の画像間距離を記録した行列

距離が小さくなるように設計したものである．

最初の for 構文処理では，画像間距離の近い画像の組み合わせを距離の近い順に作業変数 tmpP，tmpQ に記憶していく．次の for 構文処理では，最初の for 構文処理で選択したグループ内の画像の組の距離の合計を計算する．最後の 2 重 for 構文処理は，ビデオ画像グループと参照 WWW 画像グループとの最終的なグループ間距離を求めるもので，グループ内における処理対象の画像間距離の総和を計算した後，2 番目の for 構文処理で計算した画像の組の距離合計との比を計算する．また検索されたグループにおいて最初の for 構文処理で選択された各画像の組の参照 WWW 画像を図 3[4] のステップで最終的に求める検索画像とした．

4. 実 験 結 果

本章では提案手法の性能評価結果について報告する．

4·1 実験に関する指針

本研究で提案する画像検索手法はユーザが撮影したビデオ画像にアノテーションを付与するために設計した．正しいアノテーションを付与するにはビデオ画像に対応した WWW 画像をどれだけ正確に検索してこれるかが重要であるが，この検索正解率に影響を及ぼす要因として

幾つか考慮すべきことがある．

ビデオ画像に含まれる不正解画像の数

アノテーションの付与にあたっては，ユーザが各観光名所の代表的な景色やランドマーク的な建物を旅の記念に撮影しておき，ヒントとしてアノテーション処理時にシステムに提示することを想定している．この時，ユーザが観光名所の代表的な景色と思っていても参照 WWW 画像の収集に用いた WWW ページに想定する観光名所の代表的な画像がない場合もありえる．この場合 WWW ページが取り上げていない画像は不正解画像として検索正解率に悪影響を及ぼすと考えられる．

参照 WWW 画像に含まれる不正解画像の数

逆に WWW ページに想定する観光名所の代表的な画像はあるが，ユーザがシステムに提示しなかった画像も不正解画像として検索正解率に悪影響を及ぼすと考えられる．

以下の実験では，上記 2 種類の不正解画像を意図的に作成し，その数を変えながら，ユーザが提示したビデオ画像に対して，提案手法が幾つの対応する画像を検索できるかの能力を評価した．

4·2 テスト用データセット

テストのために参照 WWW 画像データとビデオ画像データを用意した．

参照 WWW 画像としては，国内外の自治体，政府観光局，旅行社など一般的な観光ガイドに関連する WWW ページから 5 つ程度の画像が収集できる観光地を選択し，観光地にある各観光名所に関する WWW ページから観光地ごとに 5 つの画像を収集し，参照 WWW 画像とした．なお，観光ガイドの WWW ページだけから 5 つの画像を収集することができない場合には，画像検索エンジンにより不足した WWW 上の画像を収集した．WWW から画像を収集した観光地は 242 地域 (グループ (G_p)) で，参照 WWW 画像は全部で 1210 枚 (= 242 地域× 5 画像) である (図 6).

検索を行うビデオ画像としては，上記 242 の観光地の中から 5 地域 (浅草，皇居，神田，柴又，上野) を選択し，実際にビデオカメラによってその地域の風景を撮影した後，撮影した各ビデオ画像からそれぞれ 8 枚の代表的な画像を選択した．この 8 枚の画像のうち 4 枚の画像は，参照 WWW 画像に対応する画像が存在するものを選択し，残りの 4 枚は対応する参照 WWW 画像がない不正解画像を意図的に混ぜた．この結果，各観光名所に対する参照 WWW 画像も 5 枚のうち最低 1 枚はビデオ画像に対応するものがない不正解画像となっている (図 6 の灰色部の画像リスト).

図 7 に，実験に用いたビデオ画像と参照 WWW 画像の対応した例を示す．図 7 に示すように，それぞれ対応する画像の間には，被写体の見かけのサイズ，撮影アングルなどに違いがある．このような違いの吸収に関しては [柳井 04] で類似画像の検索手法が検討されている．さらに，これらの画像には，紙面上では分かりずらいが，天候状態や周囲の環境による陰影など，明るさ（すなわち画質）に違いがあり，一見するよりも画像検索処理は難しくなっている．

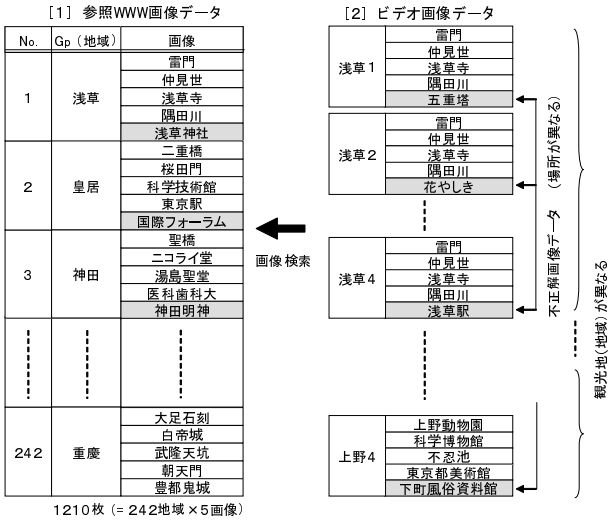


図 6 参照 WWW 画像とビデオ画像のリスト (ビデオ画像 5 枚検索の場合)

図 7 ビデオ画像と参照 WWW 画像の例

4·3 ビデオ画像に不正解画像がない場合の実験結果

初めに検索したいビデオ画像のすべてに，対応する参照 WWW 画像がある場合の検索性能を評価した．具体的には 5 地域で撮影したビデオ画像から参照 WWW 画像に対応する画像がある画像各 4 枚，合計 20 枚を取り出し，参照 WWW 画像 1210 枚の中から対応する画像が検索できるか否かの性能を調べた．

実験には色特徴，wavelet 特徴，SIFT 特徴の 3 種類の画像特徴を適用し，各特徴だけの場合，wavelet 特徴と SIFT 特徴を組み合わせた場合，およびこの特徴の組み合わせにグループ処理を適用した場合 (グループ法) の各実験結果について比較を行った．色特徴については，以下の実験結果 (図 8) で説明するように検索性能が良くなかったので，他の特徴との組み合わせた実験およびグループ処理を適用した実験は行わなかった．

図 8 は，[柳井 04] で提案されている色特徴を用いた画像検索方法による検索結果を示している (グループ処理を行っていない)．この方法は，図 1(2) で説明した画像における 25 の領域ブロックで wavelet 特徴の代わりに色特徴として $Lu*v*$ 色空間の 3 次元色ベクトルを適用して EMD を計算し画像を検索する．

横方向の位置はビデオ画像に対応し，縦方向の位置は参照 WWW 画像に対応している．○で示したデータが画像検索によって対応する画像と判定されたことを示している．対応するビデオ画像と参照 WWW 画像は，同じ順番の位置関係にあり，対角線上に○がある場合は，ビデオ画像から対応する参照 WWW 画像が正しく検索できたことを示している．色特徴のみを用いた場合は 5%(すなわち，20 枚の画像のうち 1 枚) のみ画像が正しく検索された．

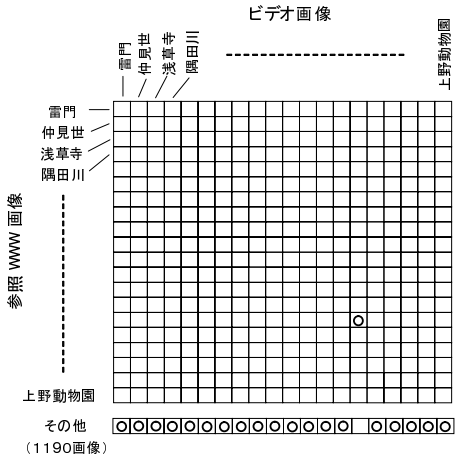


図 8 色特徴のみによる実験結果

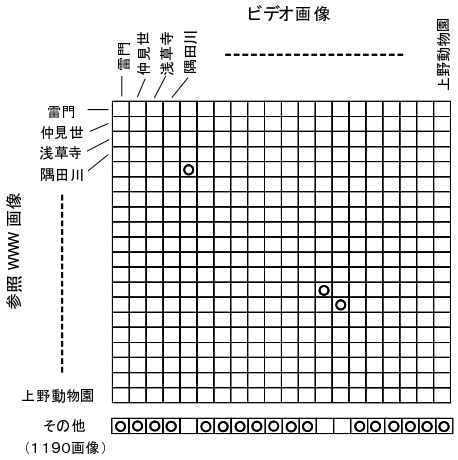


図 9 wavelet 特徴のみによる実験結果

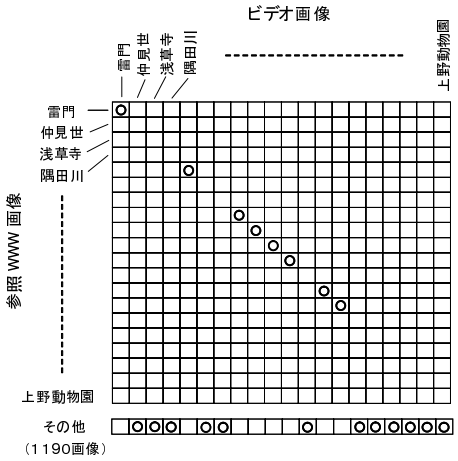


図 10 SIFT 特徴のみによる実験結果

図 9 は，wavelet 特徴のみを用いた検索結果で，15%(すなわち，20 枚の画像のうち 3 枚) の画像が正しく検索され，図 8 の色特徴の場合よりも検索性能が良いことが分かる．図 10 は，SIFT 特徴のみを用いた場合の検索結果であり，40% (すなわち，20 枚の画像のうち 8 枚) の画像が正しく検索された．この結果は図 8 の色特徴のみの場合および図 9 の wavelet 特徴のみの場合の結果より優れている．

図 11 は，wavelet 特徴と SIFT 特徴の両方を組み合わせた場合の検索結果である．両方の特徴量を用いる場合は，3·3 節で説明したように，それぞれの特徴を使って計算した画像間距離を重み係数 ($k = 0.4$，値に関する考察は 5·2 節を参照のこと) を掛けて加算したものを最終的な画像間距離として検索を行っている．この場合，45%(すなわち，20 枚の画像のうち 9 枚) の画像が正しく検索された．

図 12 は wavelet 特徴と SIFT 特徴の両方を利用し，画像のグループ処理を適用した場合 (グループ法) の結果を示している．灰色部が対応する画像と判定されたことを示している．すなわち 20 枚の画像中で 18 枚の画像が正しく検出されており，画像検出の正解率は 90%であった．またこの時，すべてのビデオ画像グループに対して対応する参照 WWW 画像グループを正しく選択することができ，誤りは同一グループ内の不正解画像によるものであった．

表 1 は，それぞれの画像特徴を適用した場合の検出性能を比較したものである．この結果から，提案手法は画像検出率を，従来手法で最も良好な SIFT 特徴のみの場合の 40%から 90%に改善していることが分かる．

表 1 各画像特徴の場合の検索結果 (ビデオ画像が 1 グループ当り 4 枚の場合)

検索に用いる画像特徴	正解率
色特徴	5%
wavelet 特徴	15%
SIFT 特徴	40%
wavelet + SIFT 特徴	45%
wavelet + SIFT 特徴 + グループ処理 (グループ法)	90%

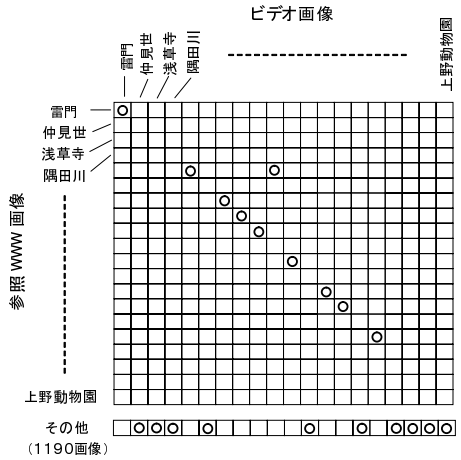


図 11 wavelet 特徴と SIFT 特徴の両方を用いた実験結果

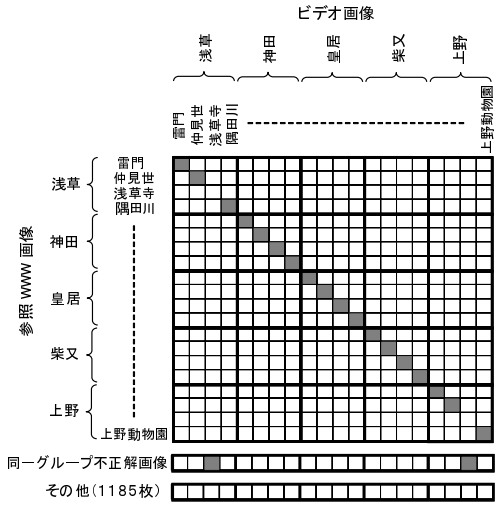


図 12 wavelet 特徴＋ SIFT 特徴＋グループ処理の実験結果

4·4 ビデオ画像の数を変え不正解画像数を変化させた場合

次にビデオ画像から選ぶ画像の数を 2 枚 ∼7 枚に変え不正解画像を増やした場合の結果を示す.

ビデオ画像から選ぶ画像の数が 2 枚〜4 枚の場合，いずれの画像も参照 WWW 画像があるものを選んで実験を行った．この場合，参照 WWW 画像に不正解画像が含まれており，ビデオ画像には不正解画像が無い場合の実験となっている．またビデオ画像から選ぶ画像の数が 2 枚の場合，参照 WWW 画像に対応する画像は 4 枚あるので，${}_4C_2 = 6$ 通りのビデオ画像の選び方がある．従って，選び方の総数 30 地域 (= 5 地域× 6 通り) として選んだビデオ画像グループについて 1210 枚の参照 WWW 画像を使って実験を行った後，結果の平均を求めた.

また，ビデオ画像から選ぶ画像の数が 5∼7 枚の場合，うち 4 枚は参照 WWW 画像にあるものを選び，残りは可能な不正解画像の組み合わせをすべて試した．例えばビデオ画像から選ぶ画像の数が 7 枚の場合，うち 3 枚は

表 2 検索する画像の組み合わせ数

検索するビデオ画像数 (n_p)	2	3	4	5	6	7
正解のある画像数 (ss)	2	3	4	4	4	4
ビデオ画像グループ中の不正解画像数	0	0	0	1	2	3
参照 WWW 画像グループ中の不正解画像数	3	2	1	1	1	1
画像の選び方の数 (sa)	6	4	1	4	6	4
画像グループ選び方総数 (sb)	30	20	5	20	30	20

参照 WWW 画像にない不正解画像のすべての組み合わせを実験した．すなわち，実質的に 4($=_4 C_3$) 通り × 5 地域の実験を行い (表 2)，結果の平均を求めた．この場合，参照 WWW 画像に不正解画像を 1 枚含んだ上で，ビデオ画像に含まれる不正解画像の枚数を変化させた実験となっている.

図 13 に，wavelet 特徴，SIFT 特徴の各特徴量のみ用いた場合，wavelet 特徴＋ SIFT 特徴を用いた場合，wavelet 特徴と SIFT 特徴の組み合わせにグループ処理を適用した場合 (グループ法) の 4 種類の結果を示す．重み係数 k の設定は，グループ処理を適用した場合 ($k = 0.4$) と適用しない場合 ($k = 0.2$)，それぞれに対して検索性能が最適となる k を設定した．*1

ビデオ画像の検索枚数 n_p によって，正解のある画像数 ss および画像の選び方の数 sa と画像グループの選び方の総数 $sb(= sa \times 5)$ が変るので，ここでは，そのすべての場合について実験を行い，正解の画像が検索できた数から平均を計算した．すなわち図 13 に示す画像検索の平均正解率 Ir は，選び方 i ($1 \leq i \leq sb$) 番目の画像グループにおいて正解の画像を検索できた数を r_i として，以下の式で計算した.

$$Ir = \left(\sum_{i=1}^{sb} r_i\right) / (ss \cdot sb) \qquad (9)$$

図に示した様に参照 WWW 画像のすべてに対応する画像があるビデオ画像数 4 枚の場合の検索性能が一番良い．また不正解画像が多くなるに従って検索性能は低下する．しかし，実験した範囲では不正解画像が多くなっても正解率は 58%を越えており，グループ処理を適用した場合の検索結果はグループ処理を適用しない何れの場合よりも良好であった.

5. 考　　察

5·1 色特徴に関する考察

前述の色特徴を使った結果 (図 8，5%) は，[柳井 04] で報告されている結果よりも悪い結果となっている．これは，ビデオ画像と参照 WWW 画像の撮影条件が異なること，撮像記録系などデバイス系に基づく色再現性の

*1 予備実験の結果、どちらも $0.2 \leq k \leq 0.4$ の範囲で良好な検索結果が得られることが分かったが，若干の性能差があるためそれぞれの場合で最適な値でグラフにした.

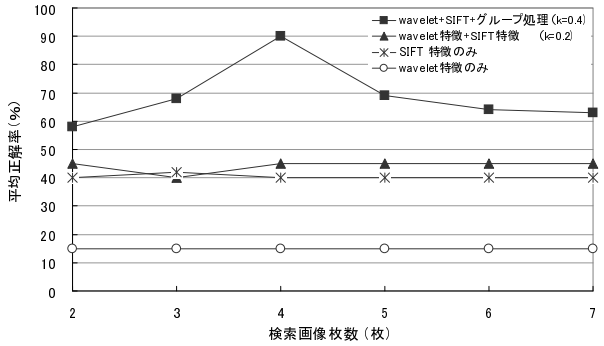


図 13 ビデオ画像の数を変えた場合の結果

特性が異なること，元の参照 WWW 画像とビデオのエンコード方式の違いなど，同一対象の画像であっても画質が異なり検索性能が出ないためと考えられる．

特に，天候や周囲の明るさなどの撮影状態によってビデオ画像の明るさや色合いが影響を受けやすいことには注意を要する．すなわちビデオ画像は明るさや色合いが校正されずに画質が異なってしまい，色特徴を用いた手法が有効に機能しなかったことが推定される．

図 14，図 15 は，このような考察を確認するために行った補足実験を説明した図である．

実験では天候が異なることによる画質の変化を模擬するために画像処理プログラムにより原画像の画質を変え，画像間距離に及ぼす影響を調べた．具体的にはあらかじめ選択したビデオ画像を基準画像 (ここでは上野：「科学博物館」の画像) として，画像処理により画像の明るさを変化させて，基準画像と画像変換後の画像との距離を計測した．図 14 においてγはガンマ補正の係数で，$\gamma = 1.0$ の時が基準画像 (原画像) を意味する．$\gamma < 1.0$ の場合は基準画像よりも画像が暗くなり，$\gamma > 1.0$ の場合は基準画像よりも画像が明るくなっている．

図 15 は，色特徴と wavelet 特徴の各特徴のみに基づく画像間距離（正規化 EMD）の変化を示すものである．変化を見るために色特徴と wavelet 特徴どちらも全く別の画像 (柴又：「寅さん記念館」の参照 WWW 画像) とγ = 1.0 の基準画像との画像間距離をそれぞれの特徴について 1 として，γ値を変えた時の基準画像と画像変換後の画像との距離を調べた．明るさが大きく変化した場合 (例えば，曇りの日に撮影した場合や快晴の日に撮影した場合など)，同一画像であっても距離が大きくなる (即ち異なる画像と判断される) ことがわかるが，色特徴の方が影響を受けやすい．

図 15 に示した結果は「ビデオ画像には明るさや色合いが校正されずに画質が異なってしまい，結果として色特徴を用いた手法が有効に機能しなかった」という考察を支持している．

SIFT 特徴も画像のスケーリング（拡大，縮小），回転，カメラの撮影アングルなどの変化の他，明るさの変化に

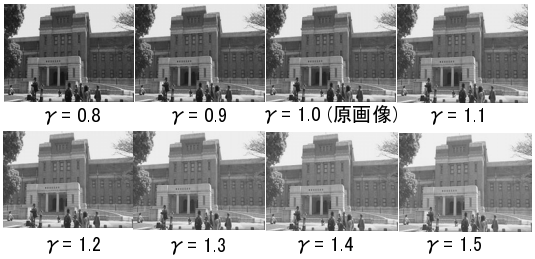


図 14 画像の明るさを変えた場合の画像例

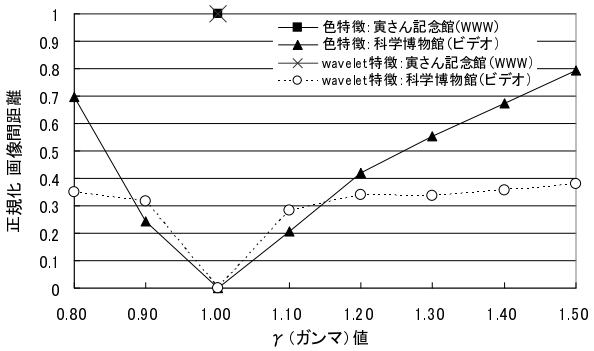


図 15 画像の明るさが画像間距離に及ぼす影響

起因する影響が少ない [Lowe 04]．明るさの変化に対して色特徴に耐性を持たせる手法 [Finlayson 05] もあるが，本論文では予備検討 [村林 07] の結果も考慮し，wavelet 特徴と SIFT 特徴を組み合わせて用いた．これは，上記 wavelet 特徴の場合と同様に明るさの変化に対する耐性を期待すると共に，明るさ以外の変化についての SIFT 特徴の耐性も期待してのアプローチである．

5·2 wavelet 特徴と SIFT 特徴の組み合わせに関する考察

提案手法で利用している画像特徴である SIFT 特徴と wavelet 特徴を個々に比べた場合，SIFT 特徴を用いた方が良好な結果を得ることが多かったが，逆の場合も存在した．図 16 はそのような画像の例であり，画像による SIFT 特徴を計算する上で重要な keypoint 対応数の違いを示している．本研究で想定する検索対象の画像には，図 16(2) のように SIFT 特徴では十分な数の特徴を抽出できない画像もある．本論文の提案手法では wavelet 特徴と SIFT 特徴を組み合わせて利用することで，検索性能を向上させている．

図 17 に重み係数 k を変化させた時の正解率の変化を示す．図から分かるように，ビデオ画像枚数に応じて最適な重み係数 k は異なるが，前節の結果では，検索枚数が 4 枚における最良の結果をもたらす係数 k(具体的にはグループ法では $k = 0.4$) を採用した．$k = 0.4$ と $k = 1.0$ の正解率の差は，SIFT 特徴では画像間距離が不正確になる場合に，wavelet 特徴に基づく補正が働いたことを示

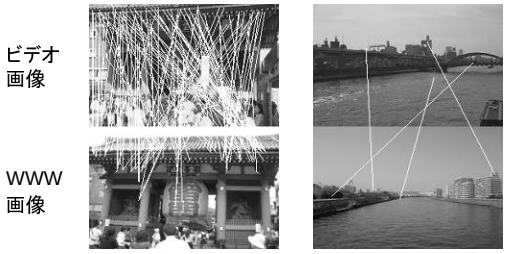


図 16 画像による keypoint 対応数の違いの例

表 3 同一グループ内における WWW 画像間距離の例

参照 WWW 画像 （k= 0.4）

参照 WWW 画像	雷門	仲見世	浅草寺	隅田川	浅草神社
雷門	0	0.99	0.96	0.91	0.88
仲見世	0.99	0	0.97	0.98	0.94
浅草寺	0.98	0.98	0	0.99	0.99
隅田川	0.95	1.0	1.0	0	0.92
浅草神社	0.85	0.90	0.93	0.84	0

している．グループ処理を用いなかった場合の正解率は，SIFT 特徴に wavelet 特徴を組み合わせても，SIFT 特徴単独の場合よりも大きくは改善されなかった (表 1) が，グループ処理を組み合わせた場合の効果は大きかったことは注意を要する．

5·3 グループ内の画像間距離の相互関係について

表 3 に一つのグループ内における各画像間距離の例を示す．本研究では各検索対象の画像は，それぞれが似ていない画像であることを想定している．これにより，前記 3·4 節で説明したグループ処理が良好に機能する．幾つかの参照 WWW 画像グループを見た範囲では表 3 に示したようにグループ内における各画像間距離は離れており，提案手法が有効に働く理由の一つになっていると考えられる．

5·4 今 後 の 課 題

今回提案手法を評価するにあたって，不正解画像を意図的に作成し，その数を変えながら実験を行った．これは「ユーザが観光名所の代表的な景色と思っていても参照 WWW 画像の収集に用いた WWW ページに想定する観光名所の代表的な画像がない」場合や「WWW ページに想定する観光名所の代表的な画像はあるが，ユーザがシステムに提示しなかった」場合の提案手法の画像検索性能を評価するために行ったものである．表 2 に示した様に，不正解画像の数を変えた時には，すべての考えられる組み合わせの平均を計算するなど，極力恣意性を排除する実験を行ったが,「ユーザが観光名所の代表的な景色として適切な画像を選べるか？」「その場合の不正解画像の数は実験を行った範囲内に収まっているのか？」といった評価は行っていない．

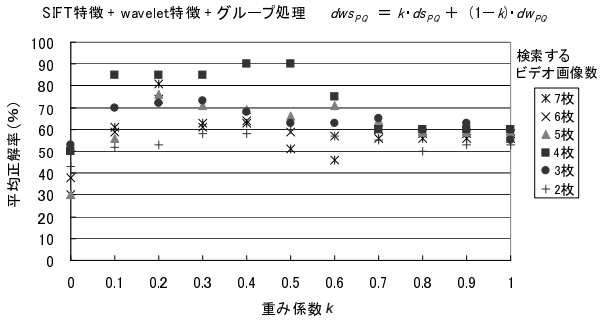


図 17 SIFT と wavelet 特徴の重み係数の画像検索特性

またグループ法は重み係数 k によって性能が異なる (図 17)．SIFT 特徴のみの場合と wavelet 特徴＋ SIFT 特徴の場合では正解率に大きな差はないが (表 1)，wavelet 特徴＋ SIFT 特徴にグループ処理を適用すると性能が大きく改善されている．この性能改善に対する理論的な解析は，今回の実験結果だけからは十分にはできていない．

上記 2 つについて考察を深めることと，提案手法を適用したアノテーションシステムを実装し実際の環境でその評価を行うことは今後の課題である．

6. ま と め

本論文では，ビデオ画像へのアノテーションのための画像グループに基づいた画像検索手法 (グループ法) について提案した．提案した手法の特徴は，以下の 2 点である．

(1) wavelet 特徴と SIFT 特徴を組み合わせて個々の画像間距離を評価する．
(2) 事前に検索対象の画像を関係するものごとにグループ化しておき，グループ間の距離も利用して類似画像の選択を行う．

更に，WWW から収集した参照 WWW 画像とビデオカメラを使用して撮影したビデオ画像を用いた実験で，次のことを明らかにした．

(1) ビデオ画像と WWW 画像のように撮影条件が異なる画像の検索でも，wavelet 特徴と SIFT 特徴を組み合わせグループ処理を適用することで高い検索性能を出すことができる．
(2) ビデオ画像と WWW 画像の間の類似画像検索において，ビデオ画像に不正解画像が無い場合の検索性能を，提案手法を用いない場合の 40%から 90%に，不正解画像が存在する場合でも 58%に改善することができる．

提案手法を適用したアノテーションシステムを実装し，その評価を行うことが今後の課題である．

付録: EMD 計算について

付録として (5) 式におけるビデオ画像 P における番号 m の領域ブロックと，参照 WWW 画像 Q における番号

m の領域ブロックの EMD(P_m, Q_m) の計算方法を示す．これは文献 [Rubner 98] によるものである．

まず，それぞれの領域ブロックの画像の特徴 (文献 [Rubner 98] では signature と呼ぶ) を次式の通りとする．*2

$$P_m = \{(\mathbf{p}_1, w_{p1}), ..., (\mathbf{p}_{15}, w_{p15})\} \quad (10)$$

$$Q_m = \{(\mathbf{q}_1, w_{q1}), ..., (\mathbf{q}_{15}, w_{q15})\} \quad (11)$$

$$(1 \le m \le 25)$$

ground distance g_{ij} を次式の通りとする．

$$g_{ij} = |\mathbf{p}_i - \mathbf{q}_j| \quad (12)$$

$$(1 \le i \le 15, \quad 1 \le j \le 15)$$

P_m を供給地，Q_m を需要地，g_{ij} を供給地から需要地までの輸送コストとする輸送問題を考え，$\sum_{i=1}^{15}\sum_{j=1}^{15} f_{ij} g_{ij}$ の最小値を与える f_{ij} を下記の条件のもとに求める．

$$f_{ij} \ge 0 \quad (1 \le i \le 15,\ 1 \le j \le 15) \quad (13)$$

$$\sum_{j=1}^{15} f_{ij} \le w_{pi} \quad (1 \le i \le 15) \quad (14)$$

$$\sum_{i=1}^{15} f_{ij} \le w_{qj} \quad (1 \le j \le 15) \quad (15)$$

$$\sum_{i=1}^{15}\sum_{j=1}^{15} f_{ij} = \min(\sum_{i=1}^{15} w_{pi},\ \sum_{j=1}^{15} w_{qi}) \quad (16)$$

この時 EMD(P_m, Q_m) を次式の通りとする．

$$\mathrm{EMD}(P_m, Q_m) = \frac{\sum_{i=1}^{15}\sum_{j=1}^{15} f_{ij}\, g_{ij}}{\sum_{i=1}^{15}\sum_{j=1}^{15} f_{ij}} \quad (17)$$

◇ 参 考 文 献 ◇

[Abowd 03] Abowd, G. D., Gauger, M., and Lachenmann, A.: The Family Video Archive: An annotation and browsing environment for home movies, in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval MIR '03*, pp. 1–8 (2003)

[Finlayson 05] Finlayson, G., Hordley, S., Schaefer, G., and Tian, G. Y.: Illuminant and device invariant colour using histogram equalisation, *Pattern Recognition*, Vol. 38, No. 2, pp. 179–190 (2005)

[Lowe 04] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004)

[Narasimhan 00] Narasimhan, S. G. and Nayar, S. K.: Chromatic Framework for Vision in Bad Weather, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2000 (CVPR 2000)*, pp. 598–605 (2000)

[Rubner 98] Rubner, Y., Tomasi, C., and Guibas, L. J.: A metric for distributions with applications to image databases, in *Sixth International Conference on Computer Vision*, pp. 59–66 (1998)

[Rubner 99] Rubner, Y. and Tomasi, C.: Texture-Based Image Retrieval Without Segmentation, in *ICCV1999*, pp. 20–27 (1999)

[Song 05] Song, Y., Hua, X.-S., Dai, L.-R., and Wang, M.: Semi-automatic video annotation based on active learning with multiple complementary predictors, in *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval MIR '05*, pp. 97–103 (2005)

[Wang 97] Wang, J. Z., Wiederhold, G., Firschein, O., and Wei, S. X.: Content-based image indexing and searching using Daubechies' wavelet, *International Journal on Digital Libraries*, Vol. 1, No. 4, pp. 311–328 (1997)

[Wang 06] Wang, X.-J., Zhang, L., Jing, F., and Ma, W.-Y.: AnnoSearch: Image Auto-Annotation by Search, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006 (CVPR 2006)*, pp. 1483–1490 (2006)

[Yu 06] Yu, L., Liu, Y., and Zhang, T.: Using Example-Based Machine Translation Method For Automatic Image Annotation, in *Proceedings of the 6th World Congress on Intelligent Control and Automation*, pp. 9809 – 9812 (2006)

[村林 07] 村林 昇, 倉橋 節也, 吉田 健一：画像のグループ化に基づいた画像検索法, 信学技報, Vol. 107, No. 290(IE2007-83), pp. 7–12 (2007)

[長坂 98] 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-ll, Vol. J81-D-ll, No. 8, pp. 1831–1837 (1998)

[柳井 04] 柳井 啓司：一般画像自動分類の実現へ向けた World Wide Web からの画像知識の獲得, 人工知能学会論文誌, Vol. 19, No. 5, pp. 429–439 (2004)

〔担当委員：谷口 倫一郎〕

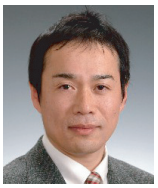
2008 年 5 月 26 日　受理

*2 以下の式の説明では記述を簡単にするため，特徴の要素データ $\mathbf{p}$, w_p，および f_{ij} g_{ij} などにおいてブロック番号 m の添え字は省略している．

著 者 紹 介

村林　昇

1982 年 東京都立大学 工学部電気工学科 卒業，1984 年 筑波大学大学院 理工学研究科 修士課程修了．同年ソニー (株) 入社．現在，筑波大学大学院　ビジネス科学研究科博士後期課程在学中．画像検索，画像処理，マルチメディア情報処理の応用などに興味を持つ．



倉橋　節也（正会員）

1995 年 放送大学教養学部 (産業と技術専攻) 卒業，1998 年 筑波大学大学院 経営政策科学研究科修士課程 (経営システム科学) 修了，2002 年 筑波大学大学院 経営政策科学研究科博士課程修了 1981〜2006 年 東京電機産業 (株) およびワイ・ディー・システム (株) 勤務．2006 年より筑波大学大学院ビジネス科学研究科 准教授．博士 (システムズ・マネジメント)．IEEE，日本ＯＲ学会，計測自動制御学会，情報処理学会などの会員．社会シミュレーション，データマイニングに興味を持つ．

吉田　健一（正会員）

1980 年 東京工業大学 理学部情報科学科 卒，同年 日立製作所 入社．1992 年 9 月 博士 (工学，大阪大学)．2002 年より 筑波大学大学院 ビジネス科学研究科教授．インターネット上の各種データを，機械学習の手法を使って解析する研究に従事．情報処理学会 会員．