Vol.2009-AVM-67 No.8
2009/12/7

映像メタデータ自動付与実現のための映像・静止画像マッチング手法の一検討

関野真洋† 青木輝勝†† 沼澤潤二††

膨大な映像コンテンツの中から希望のコンテンツを高速検索するには，映像コンテンツへのメタデータ付与が必要であるが現在メタデータ付映像コンテンツはごく一部に限られている．一方，Web ページ（動的に生成されるページは除く）上の静止画像には，検索エンジンにて自動的にメタデータが付与することが可能である．従って，映像コンテンツと Web 上の静止画像との対応付けが行えれば映像コンテンツへのメタデータ自動付与も可能となる．
本稿では，SIFT をベースとした映像コンテンツと静止画像との対応付け手法について検討した結果を報告する．

A Study on a New Image Matching Method for Automatic Metadata Generation of Video Content

Masahiro Sekino† Terumasa Aoki†† and Junji Numazawa††

Metadata for video content plays a very important role in many video related applications such as video retrieval, automatic video digest creation, copyright protection etc. However, metadata is not included in most of video content
On the other hand, metadata can be automatically given for still pictures in Web pages (except for dynamic Web pages) by Web search engines. Therefore, if video content is associated with these still pictures automatically, it also makes it possible for video content to acquire its own metadata automatically.
For this purpose, we propose a new matching method between video content and still pictures based on SIFT in this paper.

1. はじめに

近年，データの圧縮技術やネットワーク関連技術，情報ストレージ技術の進展を背景に，映像コンテンツが一般的に広く利用されてきている．しかし，ユーザが映像コンテンツの取得，蓄積を容易に行うことができる一方，映像データが大量になることで，ユーザが閲覧したいコンテンツにたどり着くことが困難になっている．

膨大な映像コンテンツの中から希望のコンテンツを高速検索するには，映像コンテンツへのメタデータ付与が必要であるが，メタデータ付映像コンテンツはごく一部に限られている．

一方，Web 上のほとんどの静止画像には，検索エンジンにて自動的にメタデータが付与されている．

従って，映像コンテンツとWeb上の静止画像との対応付けができれば映像コンテンツへのメタデータ自動付与が可能となる（図 1）．

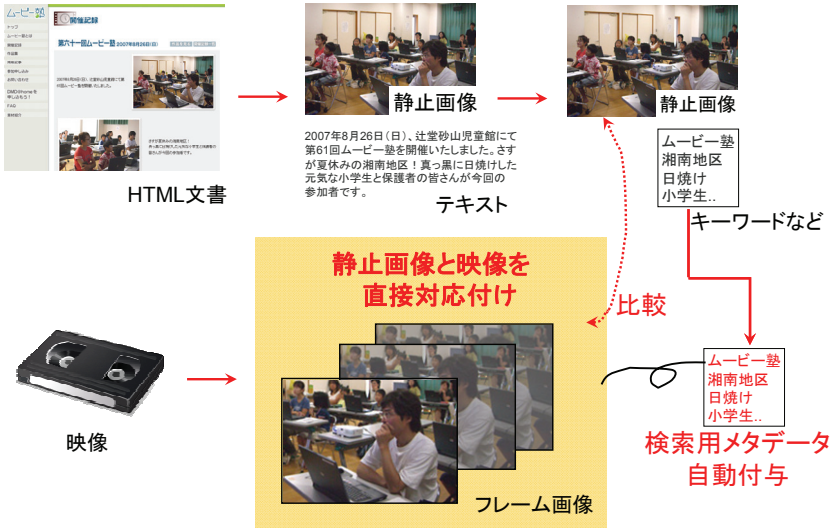


図 1　Web を利用した映像コンテンツへのメタデータ自動付与システムイメージ

† 東北大学情報科学研究科
Graduate School of Information Sciences, Tohoku University
†† 東北大学電気通信研究所
Research Institute of Electrical Communication, Tohoku University

Vol.2009-AVM-67 No.8
2009/12/7

このシステムは大きく分けて 2 つに分かれる. 1 つ目は WEB 上からキーワード検索を用いて取得した画像の特徴量を抽出する処理である.

キーワードによって WEB 上の静止画像を検索・取得し，キーワードに対応する画像特徴量を算出する. これを繰り返し行い, キーワード - 画像特徴量 DB を構築する.

2 つ目は映像コンテンツから画像特徴量を抽出し，キーワード - 画像特徴量 DB とのマッチングを行って映像コンテンツにキーワードを付与する処理である. 映像コンテンツ中のフレーム画像から画像特徴量を抽出し，先に求めた WEB 上静止画像の画像特徴量 DB とマッチングを行う. 画像特徴量間でマッチングできた場合，キーワードを映像コンテンツに付与する.

以上の提案を盛り込んだシステムを図 2に示す.

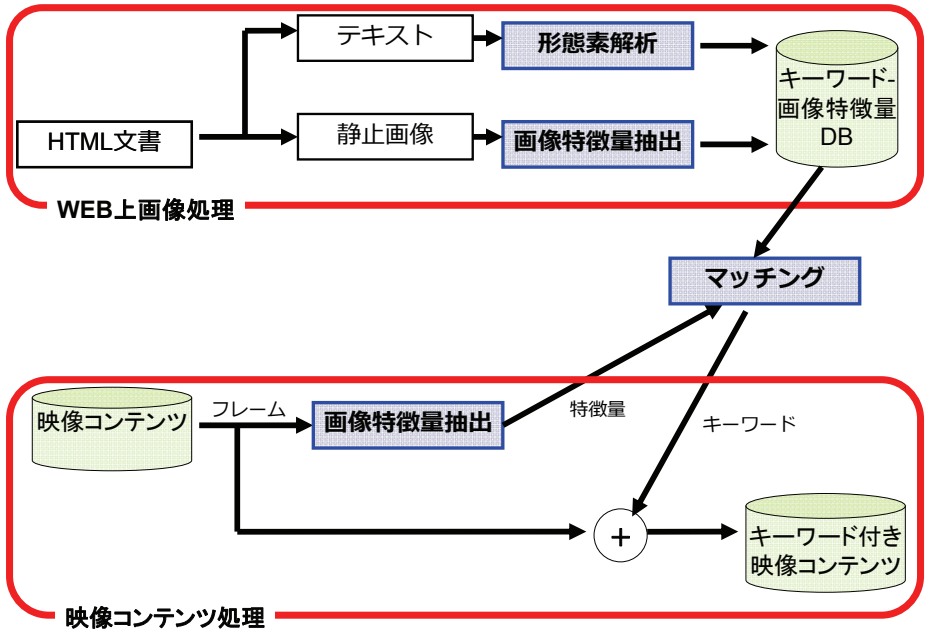


図 2　Web を利用した映像コンテンツへのメタデータ自動付与システム概念図

2. 画像特徴量抽出とマッチング

映像コンテンツと Web 上の静止画像との対応付けのため，映像コンテンツ中の最小構成単位であるフレーム画像と，Web 上の静止画像とでマッチングをとる．画像内の特徴点の対応点を探索し，特徴点中対応点が存在した割合が高いときにマッチングしたものとみなす.

画像間の対応点探索に用いる特徴量として，SIFT[1]特徴量を用いることとする.

2.1 SIFTと対応点探索

SIFT（Scale Invariant Feature Transform）特徴量は，Lowe によって提案された輝度勾配に基づく局所特徴量である．SIFT は，画像のスケール変化や回転に不変な特徴量を記述するため，画像のマッチングや物体認識に用いられる．SIFT の処理は，特徴抽出に適した点（以下キーポイント）の検出と，スケール変化・回転・照明変化に不変な特徴量を記述する 2 段階で構成されている.

2.1.1 キーポイント候補の抽出

キーポイント候補は，スケール σ の異なるガウス関数 $G(x,y,\sigma)$ と入力画像 $I(x,y)$ を畳み込んだ平滑化画像 $L(x,y,\sigma)$ の差分をとった画 $D(x,y,\sigma)$ から求める．それぞれ以下の式で求める.

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (1)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (2)$$

$$D(x,y,\sigma) = L(x,y,\sigma_{i+1}) - L(x,y,\sigma_i) \quad (3)$$

スケールの異なる平滑化画像の差分をとることを Difference-of-Gaussian（DoG）と呼ぶ．スケールを変え，複数の DoG を得ておく.

得られた DoG 画像から極値の検出を行う．極値の検出は，DoG 画像 3 枚一組で行う．極値の探索を行いたいスケールの画像のある画素に注目したとき，その周辺 8 画素，隣接する上下のスケールの DoG 画像の注目画素と周辺画素 18 画素を比較し，極値であった場合その画素を検出する．検出された画素をキーポイント候補とする.

2.1.2 キーポイントの絞込み

検出されたキーポイント候補から，キーポイントとして有効な点を選ぶ．エッジ上のキーポイントを除くため, キーポイント候補位置での 2 次元ヘッセ行列 $\boldsymbol{H}$ を計算し，主曲率を求める.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix} \quad (4)$$

ヘッセ行列 $\boldsymbol{H}$ の固有値をα，βとしたとき，対角成分の和 **Tr(*H*)**と行列式 **Det(*H*)** は次のように計算できる．

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta \quad (5)$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \quad (6)$$

さらにα=γ βとして，

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(\gamma\beta+\beta)^2}{\gamma\beta^2} = \frac{(\gamma+1)^2}{\gamma} \quad (7)$$

式（7）で得られる値は，固有値の比率のみで決まる．固有値α，βを求めず，**Tr(*H*)**，**Det(*H*)**の値から主曲率を得ることができる．この値で閾値処理することで，有効なキーポイントのみを選ぶことができる．

2.1.3 輝度勾配方向の割り当て

検出された各キーポイントに対して輝度勾配方向を割り当てる．キーポイントが検出された平滑化画像 $L(x,y,\sigma)$ の各画素の勾配 $m(x,y)$ と勾配方向 $\theta(x,y)$ を以下の式で求める．

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \quad (8)$$

$$\theta(x,y) = \tan^{-1}\frac{f_y(x,y)}{f_x(x,y)} \quad (9)$$

$$\begin{cases} f_x(x,y) = L(x+1,y) - L(x-1,y) \\ f_y(x,y) = L(x,y+1) - L(x,y-1) \end{cases} \quad (10)$$

求めた勾配の大きさ m と勾配の方向 θ から，スケール σ に比例した領域で重み付け方向ヒストグラムを作成する．作成したヒストグラムから最大値となる方向をキーポイントの輝度勾配方向として割り当てる．

2.1.4 特徴量記述

キーポイント周辺領域を2.1.3で割り当てた方向を基準とした軸に回転させる．この状態で特徴量を算出するため，回転に対する不変性を得られる．分割した周辺領域ごとの各ピクセルの輝度勾配ヒストグラムを作成する．周辺領域を小領域に分割し，それぞれの輝度勾配ヒストグラムを作成する．周辺領域を 4×4 の 16 し，それぞれにおいて 8 方向で輝度勾配ヒストグラムを作成する．これによって 4×4×8 = 128 次元のベクトルを得る．このベクトルのノルムが 1 となるように正規化し，特徴量とする．

2.1.5 対応点

対応点は次のように求める．

１．各静止画像について，SIFT 特徴量を抽出し，kd 木を構築する．

２．次に，映像中のフレーム画像から SIFT 特徴量を抽出し，キーポイントごとに類似特徴量を作成済みの kd 木より探す．類似特徴量が存在した場合，対応点とする．

特徴点の類似度は距離比を用いてを判定するものとし，1:2 未満となる場合に類似とする．

3. Web 上静止画と映像との対応付けにおける課題と提案

3.1 Web 上静止画と映像コンテンツとの対応付けにおける課題

Web 上静止画像と映像コンテンツは、言うまでもなく完全に同一ではない。しかしながら、対象物が同一である（例：同一商品のテレビ CM と一般消費者による Blog 上での評価、など）ことは頻繁に発生する。本稿ではこのような状況を前提とする。ただし、このように対象物が同一であるという状況でさえ、Web 画像と映像コンテンツとの対応付けを行うためには以下のような課題が生じる。

（１） 動画における「動きボケ」「手振れボケ」
（２） 画角の相違
（３） カメラアングルの相違
（４） 解像度の相違
（５） 画像圧縮処理の相違
（６） 照明条件の相違（フラッシュなど）

（１）は，スロー再生を目的とした撮影等以外の一般的な映像撮影の場合，各フレーム画像において動きボケや手ぶれボケが存在する可能性が高い．これは，動画は静止画の撮影とは異なり，毎秒のフレーム数が決まっておりボケを軽減するためのシャッタースピード調整が自由に選択できないためである．（２）（３）は，異なるカメラで撮影した場合に一般的に起こることであるが，撮影する範囲が異なったり，オブジェクトの見え方が異なったりすることで生じる問題である．一方の画像で見えていない領域については，対応付けができない．（４）については，自由に解像度を設定しうるために発生する問題である．まったく同じ画像であっても，解像度が異なれば，縮小された側から抽出できる SIFT 特徴点が少なくなる．（５）については，ディジタル画

Vol.2009-AVM-67 No.8
2009/12/7

像，ディジタル映像において JPEG や MPEG 等の非可逆圧縮が行われているケースがあり，これがノイズとなるものである．（6）については，静止画撮影のためのフラッシュが発生した場合，映像側の絞り調整と静止画側の絞り調整が異なるため，映像側が白とびするなどの問題である．

本稿では，（1）と（2）を対象として検討を行う．

3.2 「動きボケ」「手振れボケ」画像の問題点と対策

フレーム画像は，撮影対象の動きや手振れにより，不鮮明であることがあり，不鮮明な部分から抽出される SIFT 特徴量は対応点を持たないと考えられる．そこで，フレーム画像を縮小して解像度を落とすことで軽減することを考える．

動きボケや手振れボケは，シャッターが開いている間に移動した軌跡である．これは，周辺画素をボケ方向に移動距離分だけ平均化したものとなる．隣接画素が同一輝度値である場合には平均化しても同値をとるが，輪郭等の輝度値が急激に変わる部分の周辺における平均化がボケによる影響である．ボケの原因となった移動距離範囲において一点のみサンプリングされる状態であれば，この影響を取り除けると考えられる．したがって，フレーム画像をリサンプリングによって縮小して解像度を落とせばボケによる影響を減少させることができると考えられる．そこで，縮小済みの画像から SIFT 特徴量を抽出し，対応点探索を行う．

縮小によってスケール σ の小さい SIFT 特徴量は抽出できなくなるが，抽出できる特徴量については，スケールに対して不変な特徴量であるので，入力画像を縮小しても対応点探索は可能であると考えられる．

抽出可能な特徴点数が縮小の度合いによって異なるため，対応付けた特徴点の絶対数によって画像間の対応付けを行うことができなくなる．そこで，次式によって対応点率を定義する．

$$\text{対応点率} = \frac{N_M}{MIN\ (N_F, N_I)} \qquad (11)$$

対応点率は，フレーム画像，静止画像のどちらか全ての特徴点がもう一方の画像内のいずれかの特徴点と対応付けられたとき最大値である 1 となる．

フレーム画像と，ある静止画像とで対応点探索を行うとき，縮小後のフレーム画像に含まれるキーポイント数を N_F，静止画像に含まれるキーポイント数を N_I，対応点数を N_M としたとき，1 つのキーポイントは複数の対応点を持たないので，$N_M \leqq MIN(N_F, N_I)$ が成り立つ．N_F は縮小するほど減少するが，元画像におけるスケール σ の大きい特徴量が残ると考えられるため，N_M の減少割合は N_F の減少割合に比べて小さいと考えられる．

3.3 異なる画角の画像の問題点と対策

SIFT では．画像全体から特徴量を算出するため，特定のオブジェクト領域だけの特徴量を抽出することができない．2 つの画像中における背景領域が異なるとき，全ての特徴点が対応付けられることはない．このとき，背景部分の SIFT 特徴量はノイズとなる．

静止画像については領域指定を自動化することは困難であると考えられるが，フレーム画像については，映像からフレーム画像を切り出すときに動きベクトルの違いなどを利用することで自動的にオブジェクト領域と背景領域を分離できる．SIFT特徴量を算出するとき，検出した背景部分を領域外部分としてマスクした画像を入力し（図 3），対象領域内にのみキーポイントが発生するようにすることで，オブジェクト領域のみからSIFT特徴量を抽出できる．背景による影響を抑えることができると考えられる．

オブジェクト領域からのみ特徴点を抽出する場合，特徴点数は背景を除去しなかった場合と比べて減少しないので，3.2 同様，画像間の対応付けには式（11）の対応点率を用いる．

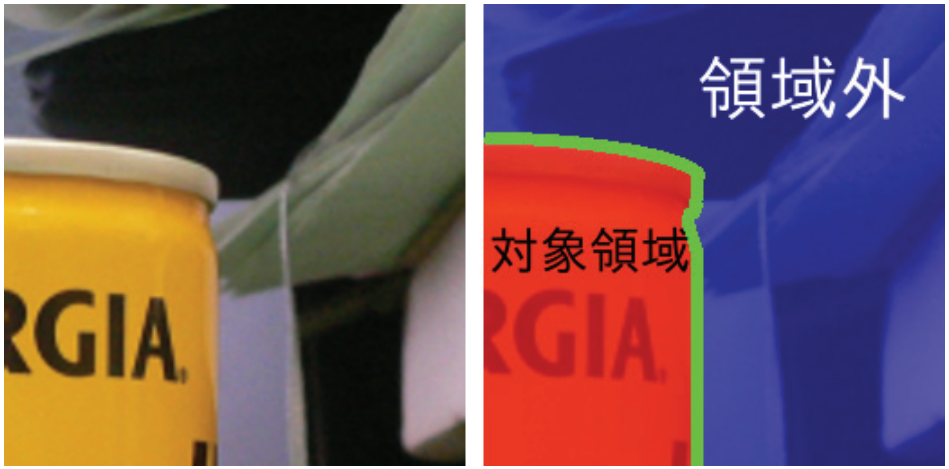


図 3 特徴量抽出対象領域の指定

4. 提案方式の評価と考察

4.1 ボケの存在する画像における縮小

背景を白色とし，実験用映像と静止画像を作成した．映像は DV で撮影し，手持ちでの水平移動を含む．静止画像はデジタルカメラで撮影し，映像中のオブジェクト（以下，対象オブジェクトという）が含まれる場合 5 例と，対象オブジェクト以外のオブ

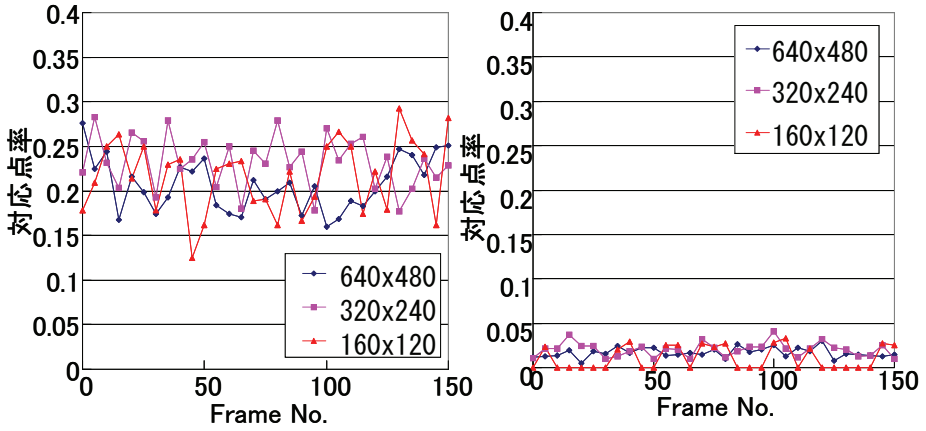
ジェクトが含まれる場合 5 例とした．なお，映像と静止画像でのオブジェクトはほぼ同一サイズになるように撮影した．

映像，静止画像ともに 640[pixel]×480[pixel]で取り込み，それらを元に 320[pixel]×240[pixel], 160[pixel]×120[pixel]に縮小したものを作成した．

静止画像を一定のサイズとし，映像を縮小したときの対応点探索，映像を一定のサイズとし，静止画像を縮小したときの対応点探索を行った．

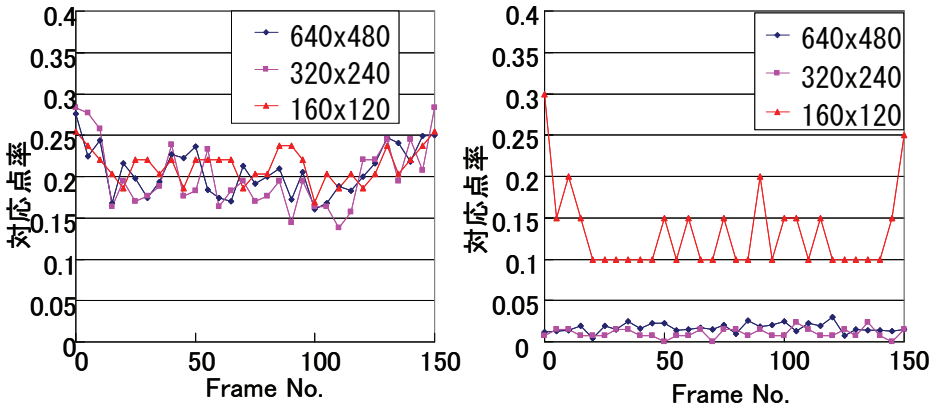
静止画像に対象オブジェクトを含む場合，映像を縮小したときに対応点率が向上した．静止画像のみを縮小した場合には変化は少ない．静止画像と対応点となるキーポイントは，低解像度となっても抽出できるキーポイントの割合が高かったと言える．

静止画像に対象オブジェクトを含まない場合，画像を 160×120 に縮小したときに急激に対応点率が上昇した（図 5b）．これは$\mathrm{MIN}(N_F, N_I)$が急激に減少したにも関わらず，N_Mが変化しなかったためである．低解像度において抽出可能なキーポイントが誤対応づけされていたものが存在したと考えられる．



(a)対象オブジェクトを含む画像　　(b)対象オブジェクト含まず

図 4 映像を各サイズに縮小したときの映像と静止画像との対応点率変化の一例



(a)対象オブジェクトを含む画像　　(b)対象オブジェクト含まず

図 5 画像を各サイズに縮小したときの映像と静止画像との対応点率変化の一例

図 6 背景が存在する画像例

4.2 背景の異なる画像における領域指定

領域指定によって対応点率が向上することを確認するため，静止画像を用いて実験

を行った．

背景は特に指定せず，実験用画像を作成した（図 6）．デジタルカメラで撮影し，領域はオブジェクト形状に合わせて手動で指定した（図 7）．画像サイズは 640[pixel]×320[pixel]とし，オブジェクトが同一で背景が異なる画像，オブジェクトも異なる画像を 5 例作成した．動きベクトル等を用いた領域指定を考慮し，対応付けに用いる 2 画像のうち，一方の画像についてマスクし，もう一方についてはマスクしないものとした．

図 7 領域指定例

対応点率の一例を図 8に示す．背景が異なり，同一オブジェクトが存在するとき，オブジェクトを除去することによって対応点率が向上することが確認された．また，全く同一の画像において，背景除去の有無のみが異なる場合には対応点率が低下することを確認した．背景によってオブジェクト領域から抽出されるSIFT特徴量に影響を及ぼすが，背景除去をしない画像同士，すなわち全く同一の画像同士では，背景によるノイズも含めて完全に一致するために対応点率は 1 となり，背景を除去した場合にはオブジェクト領域から抽出されるSIFT特徴量が異なるために 1 を下回るためであると考えられる．

映像と静止画像のマッチングを考えた場合，両者が全く同じ画像であるということは考えにくいので，背景の除去によって対応点率が低下することは起こりにくいと考えられる．

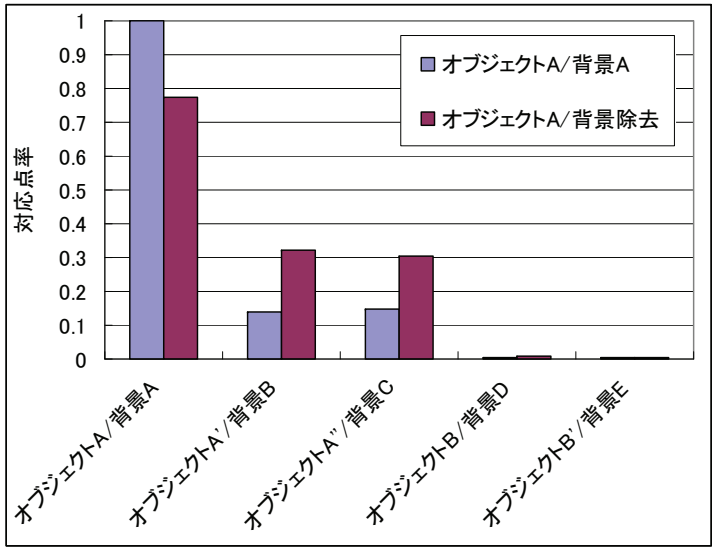


図 8 背景除去前後の対応点率の一例

5. まとめと今後の課題

Web 上静止画像に付与されているメタデータを利用して映像に対してメタデータを付与するため，映像と静止画像との対応付けを行うためのマッチング手法として SIFT を用いた基礎検討を行った．本稿では，映像中のフレーム画像は動きボケ，手振れボケ等によって不鮮明な場合があるが，縮小して解像度を低下させることで効率よく対応付けられることを示した．また，領域を指定した SIFT 特徴点抽出，対応付けを行うことにより，効率よく対応付けられることを示した．対象領域とマスク領域を自動的に指定し，対応付けを行う実証が今後の課題である．

参考文献

1) D.G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, IEEE International Journal on Computer Vision, Vol.60, no.2, pp.91-110, 2004

2) 関野真洋，青木輝勝，沼澤潤二: 映像メタデータ自動付与実現のための Web 情報を用いた画像マッチング手法の一検討，FIT2009 第 8 回情報科学技術フォーラム，I-010 (2009)