ニュース映像と Web 画像の照合によるニュースコンテンツの関連付け

奥岡 知樹† 高橋 友和† 井手 一郎†‡ 村瀬 洋†

†名古屋大学大学院情報科学研究科 〒464-8603 愛知県名古屋市千種区不老町
‡国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{okuoka,ttakahashi,ide,murase}@murase.m.is.nagoya-u.ac.jp, ‡ide@nii.ac.jp

あらまし インターネット上でニュースに関する映像・画像などのコンテンツが増加しており，これらを効率的に閲覧・検索する技術が求められている．そこで映像―画像間の照合に注目し，放送映像や動画共有 SNS に投稿されたニュース映像と Web 画像を対応付ける．これにより異なる形態の映像同士を，Web 画像との対応付けを介して関連付けることが可能となる．ニュース映像と Web 画像との対応付けは，まず映像に付随するテキスト情報を基に画像検索を行い，次に得られた Web 画像と映像中のフレームとを照合する．本研究では顔が大きく映った画像に注目し，人物周辺領域を切り出すことで，撮影角度や色情報の変化にロバストに照合することを目指す．実験により適合率 33%，再現率 50%でニュース映像と Web 画像を対応付けが可能であることを確認した．
キーワード ニュースコンテンツ，ニュース映像，Web 画像

Association of News Contents by Matching between News Video and Web Images

Tomoki OKUOKA† Tomokazu TAKAHASHI† Ichiro IDE†‡ and Hiroshi MURASE†

†Nagoya University Graduate School of Information Science Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
‡National Institute of Infomatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
E-mail: †{okuoka,ttakahashi,ide,murase}@murase.m.is.nagoya-u.ac.jp, ‡ide@nii.ac.jp

Abstract On the Internet, contents such as news videos and news photos are increasing. Techniques to browse and search these news contents efficiently are needed. We propose a method to match news videos from broadcast TV or video-sharing sites with Web images by matching videos and images. Through this process, we can associate different styles of video contents via the association. First, Web images are searched by text information associated with news videos, and then matched between a keyframe obtained from the news videos and the searched images. In this talk, focusing especially on images with a large face, we attempt to match images taken from various angles and with different color tones by comparing the adjacent regions to the faces. Through experiments, news videos and Web images were accurately associated with a precision of 33% and a recall of 50%.
Keyword News contents, News video, Web image

1. はじめに

1.1. 背景と目的

近年ブロードバンドネットワークの普及などにより，インターネット上に大量の映像や画像が存在するようになった．その中でも特にニュースに関する情報は重要であり,様々なメディアにより配信されている．例えば新聞社が開設したニュースサイトでは，ニュース記事を簡単に閲覧できるようになっており，写真付きのニュースも増えている．また，各テレビ局が開設した動画配信サイトでは，製作されたニュース映像が配信されている．さらに，ブログの中でニュースに関する意見や議論がなされており，ニュースに関する写真などをブログ上に掲載する例も多い．最近では YouTube[1]などに代表される動画共有サイトにおいて，話題となっているニュースに関する映像を投稿する例も増えてきた．このように増大しつつある画像・映像データの検索・閲覧は困難であるため，これらのマルチメディアコンテンツを効率的・効果的に利用する技術が求められている．

上記のような問題に対し，意味的または画像的に類似するマルチメディアデータを検出し，対応付けや構

[1] “YouTube”, http://jp.youtube.com/.

造化を行う研究が多くなされている[2][3]. しかしこれらは画像―画像間，映像―映像間に対するものや，それに関連したテキスト情報を利用するものであった．そこで本研究では，以下の 1.2 に挙げるような利点に着目し，映像―画像間の対応付けを手がかりとしたニュースコンテンツの関連付け手法を提案する．本講演では特に映像―画像間の対応付け手法を中心に紹介する．

1.2. 映像―画像間の対応付けの利点

映像―画像間の対応付けによって，ニュースコンテンツの閲覧，検索に関して以下の利点が考えられる．

従来は関連する映像を対応付けるために文字放送字幕（Closed Caption；CC）テキストを利用することが多かった. 一方, YouTube など動画共有 SNS での「タグ」や動画ニュース配信サイトで映像に付与された説明文などのように，インターネット上の映像に付与されるテキスト情報は様々な形態であるため，対応付けが困難である．しかし画像を介した対応付けができれば，テキスト情報の形態に影響されることなく対応付けることが可能となる．また言語を横断した検索も容易になると考えられる．

1.3. 本研究の概要

本講演ではニュース映像と Web 上に存在する画像（以下，Web 画像）の対応付け手法を紹介する．ここで，映像に付与されたテキスト情報を付随テキストとする．例えば放送映像に関しては CC テキスト，動画共有 SNS に関してはタイトルとタグなどである. 対応付けの方法として，付随テキストから生成したクエリにより画像検索エンジンを用いた検索を行い，得られた検索結果の中から映像中のいずれかのフレームと画像的な類似度が高いものを選び，映像に対応付ける．

一般的なニュース映像と Web 画像の類似度照合は，画像サイズや撮影角度の違いが存在するため，困難である．そこで本研究では問題を限定し，ニュースの話題としてインタビューや記者会見のように人物の顔が大きく映っているものに注目する．これにより，対応付けに使用する特徴として，顔の情報を積極的に用いることができる．

2. 関連研究

2.1. 画像の照合に関する関連研究

ニュース映像と Web 画像の対応付けを行う際，画像同士の類似度を評価する必要がある．画像特徴量として，Lowe による SIFT 特徴量[1]が有名である．これは対象領域の大きさや角度変化にロバストな局所特徴量であり，様々な研究でその有効性が示されている．例えば Lowe は文献[1]において，2 つの画像中の共通部分をノイズやオクルージョンに対してロバストに対応付けた．しかし人物が映っている映像・画像に注目する場合，SIFT 特徴量のような局所特徴量を用いても正しく対応が付かない場合がある．その理由を以下に示す．

（１） 光の当たり方によって，顔領域では輝度やエッジ強度及びエッジ形状が大きく変化する．

（２） ニュース写真では人物の顔に焦点を合わせるため，背景がぼける．

（３） 見栄えを良くするためや宣伝のため，人物の背景は無地に近いテクスチャや宣伝ロゴによる周期的なテクスチャであることが多い．

以上のことから，SIFT 特徴量を用いても，画像の類似度を正しく評価するのは困難であることが予想される．

2.2. ニュースコンテンツ等に関する関連研究

ニュース映像同士の対応付けに関してなされた研究は多い．これらの研究の中では人物の顔に注目した研究が多く，井手らは顔と背景を分離してその画像特徴を比較することにより，場面推定を行った[2]．小笠原らは，映像中の顔を検出し部分空間法による照合を行うことで人物の名寄せを実現した[3].

映像と画像の対応付けを行おうとした研究は少ないが，ニュース映像と新聞記事を対応付ける研究はなされており，渡辺らによる研究[4]などが挙げられる．これは，ニュース映像中のテロップと電子新聞（当時のインターネットニュース）との間で，名詞の出現頻度を計算することにより対応付けを行っている．また Fan らは SIFT 特徴量の対応を手がかりとした射影変換を用いて，講義映像とプレゼンテーションのスライドとの対応付けを行った[5]．これも映像と画像の対応付けの一種であると考えられる．

3. ニュース映像と Web 画像の対応付け

3.1. 処理の流れ

処理の流れを図 1 に示す．ニュース映像とその付随テキストを入力とする．ニュース映像に対しては，映像をショット分割し，キーフレームを選択する．本研究では，顔領域が検出され，かつ各ショット中で顔に関する画像情報（撮影角度，輝度など）が最も平均的なものをキーフレームとした．ニュース映像に対するこのような準備処理の出力は，キーフレーム及び各キーフレームの顔領域である（キーフレーム選択の際に顔領域が得られる）. 付随テキストに対しては，付随テキストを基に Web 画像を収集し，顔検出を行う．Web 画像の収集として，付随テキストから検索クエリを生成し画像検索を行った．付随テキストに対するこのような準備処理の出力は，Web 画像と各画像の顔領域である．人物周辺領域の照合では，まず顔領域の情報を

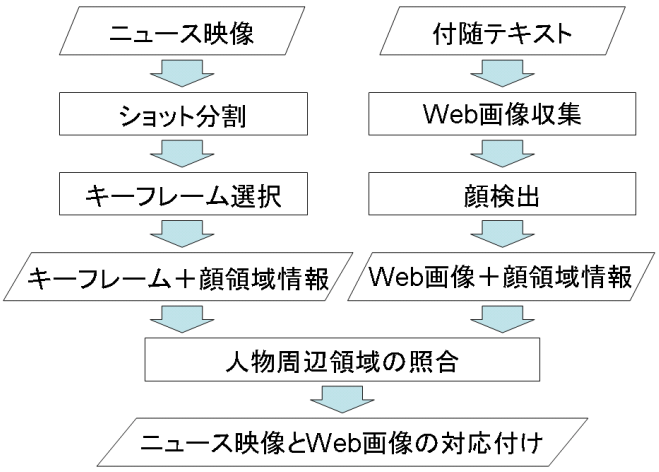

図 1　処理の流れ

基に画像サイズと輝度の正規化を行う．その後人物周辺領域を切り出し，色特徴による類似度を用いて照合を行う．この際に高い類似度を示した Web 画像を，元のニュース映像と対応付ける．

3.2. ニュース映像に対する準備処理

3.2.1. ショット分割

一般に 1 本の映像中に複数の人物が登場する可能性があり，人物ごとに固有顔を生成する必要がある．そのため，映像をまずショット分割する．ショットとは画像的に連続するフレーム群である．そのため一般に同じショット内において連続する 2 フレームの画像特徴量は極めて類似しており，フレームを時系列に沿って見て前後の画像特徴量が急激に変化する箇所をショットの切換点（すなわちカット）として検出する．以下の実験では岩成らが提案した手法[6]を用いた．

3.2.2. キーフレーム選択

映像―画像間の照合の際，映像中の全てのフレームと画像を照合するのは現実的でないため，各ショットから予めキーフレームを選択しておき,照合に用いる．照合に使用するフレームは，顔領域が検出され，かつ各ショット中で顔に関する画像情報（撮影角度，輝度など）が最も平均的なものが良いと考え，そのようなフレームをキーフレームとした．キーフレームは部分空間法を用いた顔照合により選択する．具体的には各ショットにおける固有顔を生成し，その固有顔と元のショット中の各フレームとの間で顔照合を行い，最も類似度が高くなるフレームを選択する．

3.3. 付随テキストに対する準備処理

3.3.1. Web 画像収集

Web 画像収集は，付随テキストから生成した検索クエリを基に,既存の画像検索エンジンを使用して行う．検索クエリは以下の手順により決定する．なお検索クエリは常に（人名，人名以外の名詞）のペアを 3 組用意する．

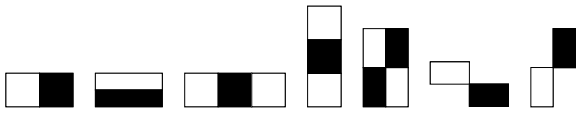

図 2　基本的な Haar-like 特徴のセット

（1） 付随テキストを形態素解析する．
（2） 最も出現回数が多い人物名詞をクエリのペアの 1 つ目とする．
（3） 人名を含まない名詞の中で出現回数が多い上位 3 つをそれぞれクエリのペアの 2 つ目とする．

なお数字や付属語などはクエリとしない．出現回数が同じものがあり，クエリのペアが 4 組以上できてしまう場合は，画像検索結果の出力数が少ないもの（0 件のものを除く）から優先的に採用した．

3.3.2. 顔検出

ニュース映像と比較して Web 画像は人物がほぼ正面を向いたものが少なく，画像全体に対する顔領域のサイズが小さいことが多い．そこで，解像度に依存せず高精度な検出が可能な顔検出手法として，三田らの Joint Haar-like 特徴による顔検出手法[7]を用いた．

Haar-like 特徴とは，画像における特徴量として，照明条件の変動やノイズの影響を受けやすい各画素の明度値をそのまま用いるのではなく，近接する 2 つの矩形領域の明度差を求めることで得られる特徴量である．図 2 に基本的な Haar-like 特徴のセットを示す．Joint Haar-like 特徴は,複数の Haar-like 特徴の共起に基づく特徴量であり，Haar-like 特徴を組み合わせることで顔の構造を表現する．

3.4. 人物周辺領域の照合

ニュース映像から得られたキーフレームと付随テキストから得られた Web 画像を照合する．ニュース映像中のフレームと Web 画像では画像サイズや撮影範囲，輝度などが異なる．そのため画像全体に対して色特徴による類似度評価を行っても，適切な対応付けは困難である．そこで，3.2 で得られたキーフレームと 3.3 で得られた Web 画像を，顔に関する情報（画像サイズ, 輝度, 画像中の顔の位置）を利用して正規化し，顔領域を含めた周辺領域を切り出して，色特徴を用いて類似度を評価する.以下に各手順について説明する．

3.4.1. 画像サイズ・輝度の補正

検出された顔領域のサイズが等しくなるように，照合する画像同士のスケールを調整する．ここではキーフレームの顔領域のサイズに合わせた．

本研究では照合する画像や映像が撮影された位置やカメラパラメータなどの環境が異なることを想定しているため,照合する画像同士が完全には一致しない．

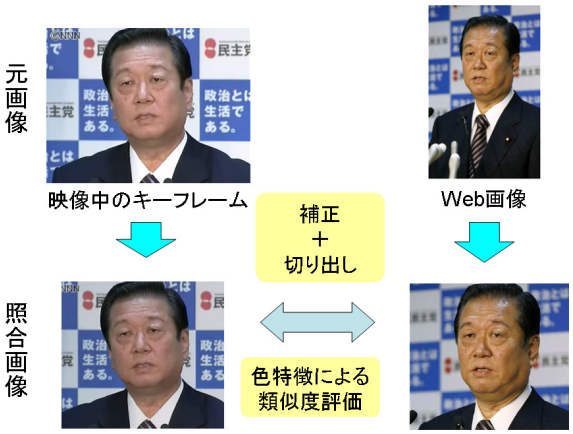

図 3　補正・切り出しの例

そのため照合する画像間で輝度や色情報などが異なることが多い．これに対処するため，顔領域が検出された画像の輝度を求め，輝度を補正する．この際，飽和を防ぐために輝度が小さい方に合わせる．

3.4.2. 人物周辺領域の切り出し

人物が映った画像中の顔周辺の矩形領域を，背景照合を行う領域として切り出す．この際切り出される領域は，顔領域に対して縦横の幅が同じ比率になるような領域のうち，照合する両画像でとれる最大のものである．ここで，人物の背景がぼけている場合や周期的なテクスチャである場合が多いために，撮影角度が若干異なっても色特徴があまり変わらないと考える．実際の画像に 3.4.1 と 3.4.2 の操作を適用した例を図 3 に示す．

3.4.3. 照合に用いる類似度

背景照合を行う際，色ヒストグラムと色コリログラム[8]の 2 通りの色特徴量を用いた．色ヒストグラムは画像中のピクセル毎の色の出現頻度の分布である．また色コリログラムは画像中の一定距離離れたピクセル間の色の組合せの出現頻度の分布である．色コリログラムは色特徴に加えて局所的な構造的特徴を表現することができる．ヒストグラム同士及びコリログラム同士の類似度はヒストグラムインターセクションにより評価する．

4. 実験と考察

実際にニュース映像と Web 画像の対応付けを行った．この際，抽出する画像特徴量を変化させ，その特性を評価する．その後，人物周辺領域切り出しによる照合を導入することによる有効性を示すために，比較実験について述べる．最後に参考実験として，手法の途中（画像検索，顔検出）で得られたデータを示す．

なお対応付けの正解は，目視により類似していると判断したものを選択した．

4.1. 使用するデータ及び実験条件

放送映像（NHK News7）及び動画共有 SNS（YouTube）に投稿されたニュース映像を 3 本ずつ使用した．放送映像に関しては 1 つの話題に関して目視で切り出して使用した．付随テキストとして，放送映像に関しては CC テキスト，YouTube に投稿された動画（以下 YouTube 動画）に関してはタイトル，タグ及び説明文を用いた．

本実験では，画像検索エンジンとして Google イメージ検索[2] と Yahoo!JAPAN[3] の画像検索を用い，それらの検索結果を合わせたものを用いた．形態素解析ソフトウェアとして MeCab[4] を用いた．キーフレーム選択時に使用する顔認識には東芝顔認識ソフトウェアを，顔検出には東芝顔検出 SDK Ver.2.0 を用いた．

4.2. ニュース映像と Web 画像の対応付け実験

4.2.1. 実験内容

提案手法により，ニュース映像と Web 画像の対応付けを行った．この際，色情報に使用する表色系を 2 種類，色特徴量として色ヒストグラムと色コリログラムの 2 種類，合計 4 種類の特徴量を抽出し，精度を評価した．色情報に使用する表色系に関して，RGB 表色系と HSV 表色系を用いた．色ヒストグラムのビンは色空間を線形に 64 分割したものを用いた．色コリログラムのビンは色空間を線形に 27 分割し，距離を 4 通りに分けたものを用いた．ビン数は 27×27×4 = 2,916 となる．しきい値は，RGB 表色系の場合は 0.7，HSV 表色系の場合は 0.5 とした．これらは経験的に決定した．

4.2.2. 実験結果

図 4 に対応付けの例を示す．図 4 の上方に示したようなフレームが存在するニュース映像に対し，対応付けられた Web 画像を，正解データと不正解データに分類して右に示す．また表 3 に 4 種類の画像特徴量を用いて対応付けを行った際の適合率，再現率及び F 値を示す．なお表に示される値は，放送映像，YouTube 動画をそれぞれ 3 本ずつ使用し，それらの誤検出数，検出漏れ数などの和をとって算出されたものである．

図 4 より，ニュース映像と Web 画像の対応付けが可能であることを確認した．また表 1 より，色ヒストグラムより色コリログラムを用いた方が 5%程度の精度向上をしていることが分かる．さらに，RGB 表色系よりも HSV 表色系を用いた方が 5%程度精度の向上がみられる．

---

[2] “Google イメージ検索”, http://images.google.co.jp/
[3] “Yahoo!JAPAN”, http://www.yahoo.co.jp/
[4] “MeCab”, http://mecab.sourceforge.net/

図 4　ニュース映像と Web 画像の対応付けの例

表 1　各類似度による精度の結果[%]

	適合率	再現率	F 値
RGB + ヒストグラム	25	42	31
RGB + コリログラム	30	46	36
HSV + ヒストグラム	26	59	36
HSV + コリログラム	33	50	40

4.2.3. 考察

図 4 の正解データ例 2 のように，話題が異なるが人物が同一で背景が類似している Web 画像も対応付けられたが，本実験ではこのような対応付けを正解とした．また不正解データ例 1 のように背景は類似していても人物が異なる Web 画像も対応付けられた．これは色コリログラムが色彩に関する情報に加えて，構造的な情報も考慮に入れているためであると考えられる．

HSV 表色系とコリログラムを使用した場合が最も精度が高く，適合率 33%，再現率 50%であった．

4.3. 比較実験

4.3.1. 実験内容

人物周辺領域切り出しによる照合を導入することによる有効性を示すために，比較実験を行った．比較手法では，切り出しをせずに画像全体から色特徴量を抽出して照合した．

照合の際に使用する画像特徴量やしきい値は 4.2 で示した値と同様である．

4.3.2. 実験結果

表 4 に画像全体に対して色特徴量を用いて照合を行った際の適合率，再現率及び F 値を示す．

表 2　比較手法による精度の比較[%]

	適合率	再現率	F 値
RGB + ヒストグラム	16	25	19
RGB + コリログラム	29	38	33
HSV + ヒストグラム	18	38	25
HSV + コリログラム	26	38	31

表 3　画像検索の結果数とその精度

	検索結果	正解数	適合率(%)
放送映像	907	39	4.3
YouTube 動画	882	83	9.4
合計	1789	122	6.8

表 4　顔検出された画像数とその精度

	画像数	正解数	適合率(%)
放送映像	393	37	9.4
YouTube 動画	493	81	16.4
合計	886	118	13.3

4.3.3. 考察

表 1 と表 2 を比較して，人物周辺領域を切り出すことによって対応付け精度が 5～10%向上することが分かる．また，切り出しを行わない場合でもコリログラムに関しては比較的高精度に対応付けが可能であることが分かる．

4.4. 参考実験

4.4.1. 実験内容

最後に参考実験として，画像検索で得られた画像の中に，映像中の各フレームと類似した画像が何枚含まれるか調査した．また顔検出を行った際にそれがどのように変化するかも調査した．

4.4.2. 実験結果

表 1 に，画像検索によって得られた画像数と，その中で映像中のキーフレームと類似している Web 画像数（正解数）及び適合率を示す．なお表に示される値は，放送映像，YouTube 動画それぞれ 3 本ずつ使用し，それらの検索結果数，正解数の和をとったものである．

表 2 に，顔検出された画像数と，その中で映像中のキーフレームと類似している Web 画像数（正解数）及び適合率を示す．

4.4.3. 考察

表 3 より，画像検索だけではニュース映像に関連した Web 画像を十分な精度で入手することができないことが分かる．また表 3 と表 4 を比較し，顔検出によって Web 画像が約半数に絞られることが分かる．また，正解数が若干減少している．これは，正解データであっても人物の顔が正面を向いていない，顔領域が極端

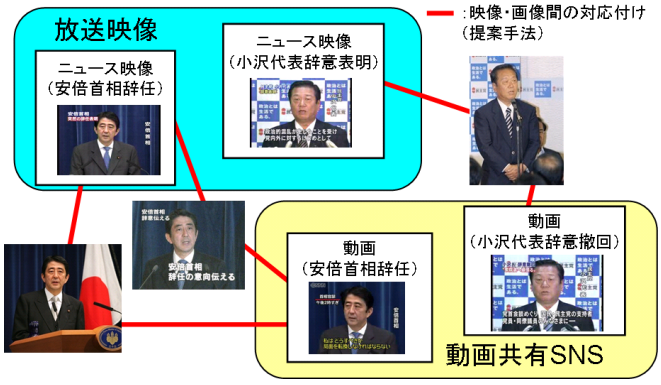

図 5　ニュースコンテンツの関連付けの例

に大きい，または小さいなどの理由で顔検出が行われなかった Web 画像が存在するためである．

5. コンテンツの関連付け

提案手法により得られた対応付けを用いて，ニュース映像と Web 画像の関連付けを行った．関連付けの例を図 5 に示す．

図 5 より，放送映像中の安倍首相の辞任に関する映像が，同じ話題に関する Web 画像と対応付けられ，さらにそれらが動画共有 SNS 中の動画と対応付けられた．また同様に放送映像中の小沢党首の「辞意表明」に関するニュース映像が，動画共有 SNS 中の小沢代表の「辞意撤回」に関する動画と関連付けられた．このように画像特徴を手がかりとして関連付けることで，テキスト情報では得にくい関連が発見されることを確認した．

このように，放送映像や YouTube といった異なる形態のテキスト情報が付与されたもの同士を，画像による対応付けを経由して関連付けることが可能になった．対応付けられた Web 画像は，ニュースサイトで公開される記事に付与された写真だけでなく，ブログや各政党が開設したホームページに存在するものなど，多岐に渡った．これにより，映像による対応付けを経由して様々な Web サイトを横断的に閲覧することが可能になる．

6. むすび

本研究では映像―画像間の対応付けを手がかりとしたニュースコンテンツの関連付けを目指し，ニュース映像と Web 画像を対応付ける手法を提案した．実験により，付随テキストによる画像検索を用いて，ニュース映像と Web 画像の対応付けが可能であることを確認した．また人物周辺領域の切り出しによって色特徴を用いた対応付けの精度が向上することを示した．さらに放送映像と動画共有 SNS のように付与されているテキスト情報の形態が異なる映像間の対応付けにも成功した．

今後の課題としてはまず，顔が映っていないものも含めた一般的なニュース映像や Web 画像に対する適用が挙げられる．この場合は顔領域を参考にした切り出しなどの操作が必ずしも行えず，本講演で述べた手法とは異なる対応付け技術が必要となる．また画像検索エンジンが提供する検索結果は時々刻々変化し，消去されることもある．そこでニュースサイトの写真付きニュースなど，ニュースに関する写真を提供するサイトに存在する画像を収集し Web 画像アーカイブとして利用することで，時間が経過しても不変な対応付けが行われると期待される．

謝　辞

実験データとして使用したニュース映像を提供して頂いた国立情報学研究所に感謝する．本研究に不可欠である技術を提供して頂いた株式会社東芝研究開発センターに感謝する．論文執筆にあたり，多くの御助言を頂いた神谷保徳先輩に感謝する．

文　献

[1] D. G. Lowe : “Distinctive Image Features from Scale-Invariant Keypoints”, Int. J. Computer Vision, Vol.60, No.2, pp.90－110 (Nov. 2004)

[2] 井手一郎，浜田玲子，坂井修一，田中英彦：“ニュース映像における人物の分離による背景の場面推定”,信学論 D-II, Vol.J84-D-II, No.8, pp.1856－1863 (Aug. 2001)

[3] I. Ide, T. Ogasawara, T. Takahashi, and H. Murase : “Name Identification of People in News Video by Face Matching”, Proc. 3rd Int. Workshop on Computer Vision meets Databases, pp.17－21 (June 2007)

[4] 渡辺靖彦，岡田至弘，角田達彦，長尾真：“TV ニュースと新聞記事の対応付け”，人工知能学会論文誌, Vol.12, No.6, pp.921－927 (Nov. 1997)

[5] Q. Fan, K. Barnard, A. Amir et al : “Matching Slides to Presentation Videos using SIFT and Scene Background Matching”, Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval, pp.239－248 (Oct. 2006)

[6] 岩成英一，有木康雄：“DCT 成分を用いたシーンのクラスタリングとカット検出”，信学技報，PRU93－119 (Jan. 1994)

[7] T. Mita, T. Kaneko, O. Hori : “Joint Haar-like Features for Face Detection”, Proc. 10th IEEE Int. Conf. on Computer Vision vol.2 pp.1619－1626 (Jan. 2005)

[8] J. Huang, S. R. Kumar, M. Mitra, and W. Zhu : “Image Indexing using Color Correlograms”, Proc. IEEE Computer Vision and Pattern Recognition 1997, pp.762－768 (June 1997)