社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

顔の部分領域別動き照合によるニュース映像からの同一場面検出

熊谷はるか† 道満 恵介††,† 井手 一郎† 出口 大輔†††,† 村瀬 洋†

† 名古屋大学大学院情報科学研究科 〒 464-8601 愛知県名古屋市千種区不老町
†† 中京大学工学部 〒 470–0393 愛知県豊田市貝津町床立 101
††† 名古屋大学情報連携統括本部 〒 464-8601 愛知県名古屋市千種区不老町
E-mail: †kumagaih@murase.m.is.nagoya-u.ac.jp, ††kdoman@sist.chukyo-u.ac.jp, †ide@is.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp, †murase@is.nagoya-u.ac.jp

あらまし 本稿では，同一場面，つまり，同一のイベントを異なる視点で撮影した同一場面をニュース映像から検出する手法の改良について報告する．従来，我々は，音声特徴と画像特徴を組み合わせた同一場面検出手法を提案してきた．しかし，音声特徴は話者と音声が一致する同一場面しか検出できないため，画像特徴のさらなる精度向上が課題であった．具体的には，従来手法では，異なる顔部品に属する特徴点軌跡が誤って対応付き，誤検出が起こるという問題があった．そこで顔領域を照合する際に，縦に 3 分割した部分領域同士を照合する改良方法を提案する．部分領域同士を照合することにより，各部分領域に含まれる顔部品同士を照合することができるため，照合精度の向上が期待できる．実際に放送されたニュース映像で同一場面検出実験を行い，提案手法の有効性を確認した．
キーワード 同一場面検出，画像特徴，部分領域，顔部品，ニュース映像

Scene Duplicate Detection from News Videos by Sub-Region-based Motion Matching of a Face

Haruka KUMAGAI†, Keisuke DOMAN††,†,

Ichiro IDE†, Daisuke DEGUCHI†††,†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University, Japan
†† School of Engineering, Chukyo University, Japan
††† Information and Communications Headquarters, Nagoya University, Japan
E-mail: †kumagaih@murase.m.is.nagoya-u.ac.jp, ††kdoman@sist.chukyo-u.ac.jp, †ide@is.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp, †murase@is.nagoya-u.ac.jp

Abstract In this report, we propose a method to detect scene duplicates from news videos. Scene duplicates are video clips taking the same event from different perspectives. To detect scene duplicates, we have previously proposed a method that combines audio features and image features. However, detectable shots with audio features were restricted. So, the accuracy with image features was needed to be improved. However, in the previous method, there was a problem that feature point trajectories of different facial parts were erroneously matched. In order to deal with this problem, the proposed method divides the face region into three sub-regions, and performs the matching between them. In the way, the same facial parts contained in each partial region are expected to be matched correctly and the accuracy should improve. Scene duplicate detection experiments were conducted using actual news videos. From the results, we confirmed the effectiveness of the proposed method.
Key words Scene duplicate detection, image feature, sub-regions, facial parts, news video

1. はじめに

近年，記憶装置や通信技術の発達により，放送映像を大量に蓄積できるようになった．放送映像の中でも，ニュース映像は実世界の出来事を記録しており，資料的観点から価値が高い．ニュース映像を資料として活用するためには，大量に蓄えた

図 1　同一場面の例

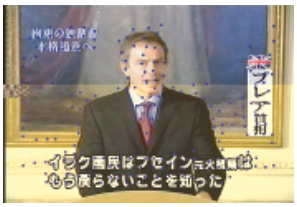
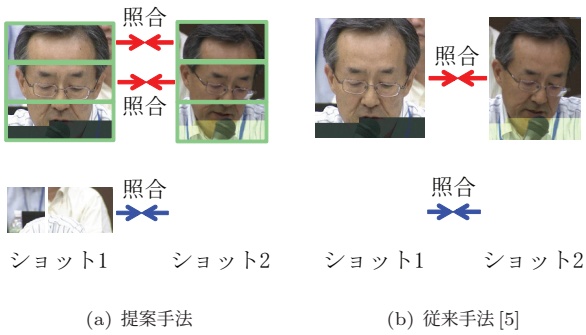


図 2　特徴抽出のための特徴点（青い点）が人物領域に少ない映像 [1]

ニュース映像を内容に基づいて扱う技術が必要となる．本研究では，そのような技術の中でも，「同一場面」をニュース映像から検出する技術に注目する．同一場面とは．図 1 に示すような，ある同一の時点・場所で起きたイベントを撮影した映像のことである．同一場面を利用すれば，異なるニュース番組で放映された，同一のイベントに関する多様な映像を，撮影視点によらず関連付け，様々な視点から視聴することができるようになる [2] [3]．よって，同一場面は，ニュース映像を内容に基づいて扱うための重要な手がかりとなる．

同一場面の検出を目的とした研究として，瀧本らは映像中のフラッシュを利用した検出手法を提案している [4]．しかしこの手法は，フラッシュによる特徴のみを用いるため，フラッシュが使われていない映像は検出できない．Wu らは，映像中の人物の特徴点の不均一性パターンに基づいた検出手法を提案している [1]．この特徴点の不均一性パターンは，人物の動きの緩急を表す特徴である．人物の動きの緩急は，撮影視点の違いに頑健であり，また，同じ撮影対象でも撮影時刻によって変化するため，撮影時刻を区別することができる．しかしこの手法は，図 2 に示すような，背景に文字や絵があるような場合（記者会見映像など）は検出が困難である．このような映像では，背景に特徴点が集まり，人物に特徴点が少なくなるため，人物の動きの緩急を正しく抽出できないためである．実際のニュース映像にはそのような状況が頻発するため，この手法をそのまま適用することはできない．また人物の動きの緩急は，ある程度の撮影視点の違いに頑健であるが，撮影視点が大きく異なる場合には検出が困難である．したがって，より高精度な検出のためには，撮影視点に影響されにくい特徴の利用が必要である．

以上の問題を解決するため，我々は音声特徴と画像特徴を併用する同一場面検出手法を提案した [5]．画像特徴による照合では，映像の各フレームを顔領域と背景領域に分割し，それぞれの領域同士を独立に照合し，その結果を統合した．顔領域の照合には既存手法と同様に特徴点の不均一性パターン [1] を利用し，背景領域の照合には色ヒストグラムを利用した．実験を行い，撮影視点が異なる映像に対する有効性を確認した．しかし，音声特徴は話者と音声が一致する同一場面しか検出できないため，画像特徴のさらなる精度向上が課題であった．具体的には，従来手法による顔領域の照合では，異なる顔部品に属する特徴点軌跡が誤って対応付き，誤検出が起こるという問題と，マイクなどにより口唇領域が隠され正確に特徴を抽出することができず，誤検出が起こるという問題があった．



(a) 提案手法　　(b) 従来手法 [5]

図 3　画像特徴による照合の概念

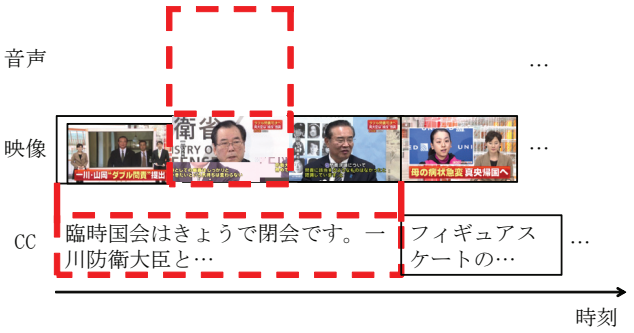


図 4　ニュース映像の一部と提案手法における入力データ（破線部分が入力データを表す）

そこで本稿では，図 3(a) に示すように，画像特徴の照合の際に，顔領域を縦に 3 分割し，部分領域間で照合するという同一場面検出手法の改良について報告する．従来手法は，図 3(b) に示すように顔領域全体間で照合した．これに対し，提案手法では部分領域間で照合することにより，各部分領域に含まれる顔部品間で照合することができる．また，照合の際にマイクなどに隠されやすい下部の領域を使用しないことで，より正確な照合を期待する．

以降，2. で提案手法について述べた後，3. で評価実験について述べ，その結果について考察する．最後に 4. で本報告をまとめる．

2. 提 案 手 法

提案手法では，ショット，音声，CC（Closed Caption; 文字放送字幕）を 1 組の入力データとし，入力データ間の照合を行う．まず，入力データについて説明する．図 4 は，ニュース映像の一部である．横軸は時刻を，映像の各サムネイル画像はショットを表す．この図に示すように，入力データとして，映像，音声，CC のそれぞれ一部分を用いる．それぞれ破線で囲まれた部分が入力データである．映像は，入力としてショットを使用する．音声は，入力としてショットに対応する区間を使

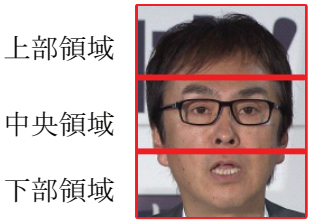


図 7　顔領域の部分領域への分割

図 8　顔領域の下部領域が隠されるショット対とその類似度

中心領域を切り出す．切り出された中心領域の中央部分を顔領域とし，左右を背景領域とする．また，背景領域はショット中で大きく変化することが少ないと仮定し，類似度計算には入力ショットの最初のフレームのみを用いる．

2.3.2　領域別の照合

領域別の照合は，人物領域の照合と，背景領域の照合からなる．以降，各処理について説明する．

a）　人物領域の照合

顔領域の照合は，Wu らによる映像中の人物の特徴点の不均一性パターンに基づく同一場面の検出手法 [1] を準用する．本研究では，図 7 に示すように，顔検出結果を縦に 3 分割した部分領域それぞれで顔領域の類似度を算出する．以降，これらの部分領域を上から上部領域，中央領域，下部領域と呼ぶ．人の顔部品の位置関係はおおよそ決まっているため，これらの部分領域間を照合することで，異なる顔部品間で動きの緩急が照合されることを防ぐ．ただし，図 8 に示すように，下部領域はマイクなどで隠されることがあるため，正確に人物の動きの緩急を抽出することが難しい．実際に，顔領域全体でこのショット対の類似度を計算すると 0.83 と高くなっており，誤検出の原因となっていることがわかる．そこで，上部領域と中央領域を使用し，下部領域は使用しない．また，Wu らの手法では，類似度の計算に特徴点軌跡の一部を用いているが，本手法では，部分領域間の類似度を計算する際，各部分領域で検出された特徴点軌跡をすべて使用する．ここでは，特徴点軌跡の検出を顔領域に限定するため，ほぼすべての特徴点軌跡が有効に働くと考えたためである．最後に，各部分領域間の類似度から顔領域の類似度を次式で計算する．S_f は入力ショット間の顔領域の類似度を表す．また，S_{f_2}, S_{f_1} はそれぞれ，中央領域の類似度と上部領域の類似度を表す．

図 9　映像の明るさが異なる同一場面例

表 1　実験に用いた映像（2012 年 6 月 7 日〜2012 年 6 月 13 日）

放送局	番組名	放送時刻
NHK 総合	ニュース	12:00 〜 12:20
東海テレビ	FNN スピーク	11:30 〜 11:55
中京テレビ	NNN ストレイトニュース	11:30 〜 11:55
CBC テレビ	JNN ニュース	11:30 〜 11:55
メ〜テレ	ANN ニュース	11:45 〜 12:00

$$S_f = \alpha S_{f_2} + (1-\alpha) S_{f_1} \tag{1}$$

ここで，α は 0 から 1 の値をとる重み係数である．

b）　背景領域の照合

背景領域の類似度として，色ヒストグラム間の距離を用いる．図 9 に示すように，ニュース映像は，録画状況や各放送局の規定によって同一場面でも映像の明るさが異なる場合がある．この明るさの違いによる検出漏れや誤検出を避けるために，色ヒストグラムには HSV 値の色相（Hue）と彩度（Saturation）を用いる．まず，領域分割されたショットの最初のフレームから色相と彩度を求め，2 次元ヒストグラムを計算する．次に，ヒストグラム間の Bhattacharyya 距離を求め，1 から減じることで背景領域の類似度 S_b を計算する．

2.3.3　個別照合結果の統合としきい値処理

顔領域の類似度 S_f と背景領域の類似度 S_b から，入力ショット間の類似度 S を次式で計算する．

$$S = \beta S_f + (1-\beta) S_b \tag{2}$$

ここで，β は 0 から 1 の値をとる重み係数である．

そして，計算された類似度にしきい値処理を適用し，入力データ対が同一場面か否かを判定する．類似度がしきい値以上であった場合は，入力データ対を同一場面と判定し，しきい値未満であった場合は同一場面でないと判定する．

3.　実験・考察

実際に放送されたニュース映像に本手法を適用することで，その有効性を評価した．

3.1　実験用映像

本実験では，2012 年 6 月 7 日から 2012 年 6 月 13 日に放送された複数の放送局のニュース映像を使用した．使用した放送局と番組，放送時刻を表 1 に示す．また，ニュース映像のフレームレートは 29.97 [frame/sec] であり，解像度は 1,440 × 1,080 [pixels] であった．これらの映像から，人物を中心に撮影されたショットを人手で計 678 ショット切り出した．全てのショットの組み合わせは ${}_{678}\mathrm{C}_2 = 229{,}503$ 組であり，そのうち同一場面対は 840 組存在した．なお，同一場面の検出精度への影響を排

図 12