映像情報メディア学会誌 Vol. 66, No. 7, pp. J251〜J259（2012）
© 2012 一般社団法人映像情報メディア学会

論 文

シーンの連続性と顔類似度に基づく動画コンテンツ中の同一人物登場シーンの同定

Identifying Scenes with the Same Person in Video Content on the Basis of Scene Continuity and Face Similarity Measurement

平井辰典†，中野倫靖††，後藤真孝††，正会員 森島繁生†

Tatsunori Hirai†, Tomoyasu Nakano††, Masataka Goto†† and Shigeo Morishima†

Abstract We present a method that can automatically annotate when and who is appearing in a video stream that is shot in an unstaged condition. Previous face recognition methods were not robust against different shooting conditions, such as those with variable lighting, face directions, and other factors, in a video stream and had difficulties identifying a person and the scenes the person appears in. To overcome such difficulties, our method groups consecutive video frames (scenes) into clusters that each have the same person's face, which we call a "facial-temporal continuum," and identifies a person by using many video frames in each cluster. In our experiments, accuracy with our method was approximately two or three times higher than a previous method that recognizes a face in each frame.

キーワード：動画像解析，マルチフレーム認識，顔認証，人物要約，3 次元顔復元，トラッキング

1. まえがき

近年，インターネット上の動画コンテンツの数が爆発的に増加し，それを楽しむユーザが増えてきている．一方，その数の多さから，ユーザが興味を持つ動画コンテンツを的確に探して閲覧することが困難な状況であるともいえる．

インターネット上のコンテンツの検索では，テキスト情報に基づく手法が広く普及している．動画コンテンツの場合でも，投稿者や視聴者が付与したタイトルや説明，タグ等のテキスト情報が検索に利用できるが，テキストとして記述される情報は限られていることが多く，映像の内容を充分に反映した検索はできなかった．テキスト情報でも，「誰が映っている何についての映像」かは判明することがあるが，動画コンテンツでは時系列情報，例えば，「いつ誰が何をしている映像」かという情報も重要となる．動画コンテンツに時系列のアノテーション情報を付与することができれば，動画中の登場人物に関する一覧性を高めてブラウジングしたり，ある人物が登場するシーンだけを検索したりすることが可能となるからである．こうした時系列アノテーション情報の自動付与は重要な課題であり，コンテンツの増加とともにそうした技術への需要は増加していくと考えられる．

本研究では，動画コンテンツに対する時系列のアノテーション情報の中でも，特にユーザが検索する上で有用かつ重要な人物名とその登場シーンに焦点を絞り，動画像の内容理解に基づいて時系列のアノテーション情報を自動付与する新しい手法を提案する．具体的には，動画像に対して顔検出を行い，検出した顔情報を動画像フレームの時間連続性に基づき伝搬させ，同一人物のクラスタを作り，クラスタ単位での認証をすることによって人物の認証精度を向上させる．本手法により，動画コンテンツへ「誰がいつ映っている映像」かという情報を自動付与できるため，ユーザは関心を持つ人物が映っている動画コンテンツを横断的に視聴することが可能となる．また，1 枚の顔画像や数フレームの動画素片を検索クエリとして，インターネット上の多数の動画コンテンツの中からその人物が写っているシーンを検索して抜き出してくることや，大量のホームビデオなどの個人の所有する動画像群に対して人物毎のインデックスを付与して整理することなどが可能となる．

2011 年 11 月 30 日受付，2012 年 4 月 11 日再受付，2012 年 5 月 23 日採録
† 早稲田大学 大学院先進理工学研究科
（〒 169-8555 新宿区大久保 3-4-1，TEL 03-5286-3510）
†† 産業技術総合研究所
（〒 305-8568 茨城県つくば市梅園 1-1-1，TEL 029-861-2130）
† Faculty of Science and Engineering, Waseda University
(3-4-1, Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan)
†† National Institute of Advanced Industrial Science and Technology (AIST)
(1-1-1, Umezono, Tsukuba-shi, Ibaraki 305-8568, Japan)

* 本研究は，JST CREST「コンテンツ共生社会のための類似度を可知化する情報環境の実現」の一環として実施されたものである．

2. 研 究 背 景

2.1 関連研究とその問題点

動画コンテンツの内容に対するアノテーションに関連した研究として，動画コンテンツに付与されるソーシャルアノテーションを利用した手法や，動画像自体を解析した手法がある．前者のソーシャルアノテーションを利用した手法では，佃らは，視聴者によって動画像の時系列に同期して付与されるコメント情報に基づき，動画コンテンツ中の登場人物毎の盛り上がり箇所等を推定し，動画コンテンツに対する時系列アノテーションとして検索に活用している[1]．しかし，ソーシャルアノテーションには主観的情報も含まれており，必ずしも動画コンテンツの内容を反映しているとは限らない．そこで，動画像自体を解析して内容を理解することで，ソーシャルアノテーションによる情報を客観的に補うことが有効である．

そうした動画像の解析による動画コンテンツの内容理解に関する研究は多く行われている．その中に TRECVID[2]（TREC Video Retrieval Evaluation）と呼ばれる，動画検索技術に関する国際的な評価ワークショップが存在する．そこで成果を挙げている技術の一つにマルチフレーム認識がある．従来では，映像内容を理解するために単一のフレームのみを扱っていたのに対し，マルチフレーム認識では，複数のフレームを扱うことで精度の向上を図っている．樋爪らは映像特徴を Bag-of-Features で表現し，マルチフレームを MKL-SVN で学習・分類したところ，キーフレームのみで認識したときに比べて大幅に性能向上し，TRECVID2010 の実験データにおいて，TRECVID2010 全チームの平均値を全クラスで上回った[3]．TRECVID では，物体認識を元に動画コンテンツの内容理解を行う課題が主に扱われており，マルチフレーム認識は人物認識でなく，一般物体認識に対して用いられていた．本研究で対象とする人物名のアノテーションや検索の目的では，映っている物よりも人物の方が重要であり，そうした従来技術がそのままでは利用できない．

一方，動画像中の顔を認識（認証）する研究も多く行われている[4][5]．しかし撮影条件は認証精度を大きく左右するため，従来の研究では，高精度の認識率（認証率）を達成するために使用用途や撮影条件を限っていることが多かった．そのため，人物の顔が多様な撮影条件下で登場する用途には利用できない．例えば，従来は同一撮影条件下の動画像から取得できる複数の顔情報を元に，一つの顔モデルを作成している研究が多く，本研究で必要な多様な撮影条件下で顔認証をするには，それぞれの撮影条件に合わせたモデルの構築が必要で現実的ではない．

2.2 提案手法による解決法

動画コンテンツにおいて，人物顔が映っているフレーム（以降，顔フレームと呼ぶ）では，あらゆる撮影条件が考えられ，そのすべてを扱うことは現実的に難しい．既存の顔認証手法でも，条件を統一するための補正や正規化の処理を加えたり，撮影条件の違いに対して頑健な特徴量を用いたりして対処する方法は提案されている．しかし，動画コンテンツにおける顔フレームでは，通常の顔認証で直面する照明や顔向き，表情などの違いの他に，オクルージョンや顔の経年変化，映像の解像度の違いなどもすべて同時に考慮しなければならならず，難しい．そのため，事前に用意した画像と照合するような顔認証のアプローチで対処するには限界がある．

顔認証において，撮影条件は認証精度を大きく左右する最も重要な要素の一つであり，高精度な顔認証を行うためには，正解データとなる顔画像に近い撮影条件下での顔フレームが必要となる．しかし，多様な撮影条件下で撮影された実動画中には，必ずしも条件に合ったデータが含まれているとは限らない．そこで我々は，フレームの時間連続性に着目し，同一人物の様々な顔フレームを蓄積することで，広範な撮影条件で撮影された動画像に対しても有効な人物顔のマルチフレーム認識の手法を考案した．具体的には，フレームの時間連続性を用い，既存の顔検出手法で検出できた顔領域情報を前後にトラッキングさせながら伝搬させていくことで，単一フレームでは顔検出ができないようなフレームに対しても顔検出を実現させた（3.3 節で説明）．そのようにして検出した顔フレームを，フレームの時間連続性を利用してクラスタリングし，さらのそのクラスタ同士を顔類似度に基づきクラスタリングすることで，同一顔情報の蓄積を行った（3.5 節で説明）．

さらに，その過程で生成される人物顔のマルチフレーム情報を元に，インターネット上の膨大な顔情報の中から人物を特定するための手法についても検討した．マルチフレームをクエリとした顔認証を行うことで，マルチフレームのうちの 1 フレームでも本人と一致していれば，該当フレームの含まれるマルチフレームすべてに対して認証したい本人の名前（以降，正解ラベルと呼ぶ）を与えることができる（3.6 節で説明）．

本手法では，顔検出やトラッキングに関する既存手法を組合せることで，動画像中の顔認証精度の向上を図っている．

3. シーンの連続性と顔類似度に基づく動画コンテンツ中の同一人物の同定手法

本研究では，まず動画像中の人物の顔を検出し，検出した顔情報を元に，前後にトラッキングすることで顔フレームを自動推定し，続いて顔認証によってその人物名を同定する手法を提案する．ここで，本手法の特長は，単に顔フレームを推定してその個々のフレーム毎に人物名を同定するのではなく，まずは顔フレームだけを推定してそれを集めた後，その中でも条件が良い数フレームにおける顔領域の画像（以降，顔画像と呼ぶ）のみを使って顔認証する点にある．本論文では，このような同一人物の顔画像群を「顔時間連続体」と呼び（**図 1** 参照），これによって様々な撮

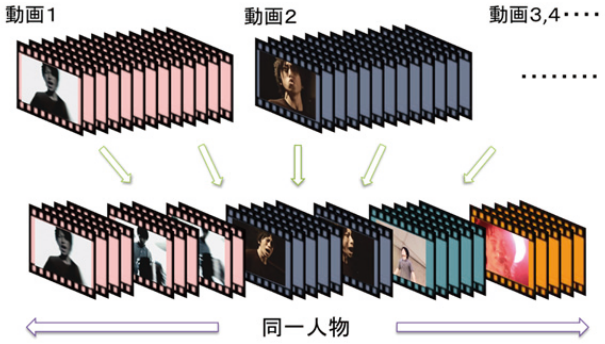


図 1　顔時間連続体の概念図
The image of facial-temporal continuum.

影条件の違いを吸収して，通常の顔認証では困難な条件にも対応した顔認証を実現することを狙う．

多くの動画コンテンツでは，認識したい対象が常に動いている．したがって，それがたとえ同一人物であったとしても，すべてのフレームが顔認証に適しているとは限らず，逆に，顔認証には条件の悪いサンプルであることも多い．本手法では，動画像フレームの時間連続性に注目することで動画像中の顔領域を検出する．フレームの時間連続性を考慮すると，あるフレームで顔が検出できたときにその前後のフレームでも同一人物の顔が，ほぼ同じ座標に存在する可能性が高い．これを利用して，動画像中の顔検出が成功したフレームの情報を元に，顔領域に関する情報をフレーム間に伝搬させながら探索することで，動画像中の直接検出することが困難な顔領域も抽出できる可能性がある．また顔認証のステップにおいて，単一フレームでは認証が困難な顔フレームに対しても，顔時間連続体の中の 1 フレームでも認証したい本人の顔画像（以降，正解顔画像と呼ぶ）と同一人物であるといえれば，膨大な数の同一人物の顔フレームすべてに対して一度に正解ラベルを付与することができる．このように提案手法は，複数のフレームを扱うことで精度の向上を図るマルチフレーム認識の一種として考えられる．

図 2 に提案手法の処理の流れの概要を示す．はじめに「ショット検出」（ショットの意味は 3.1 で後述）によって動画コンテンツをフレームの連続性が保たれる最小単位に分割する（図 2 ①）．次に，各ショットに対して顔検出を行い，「顔が検出されたフレーム」を比較して前後に顔トラッキングを行う（図 2 ②，③）．その後，検出された顔領域の縦横サイズおよび顔向きを正規化し，後述する顔に関する特徴量を元に算出した顔の類似度を元に顔時間連続体を構築する（図 2 ④，⑤）．最後に，インターネット上の顔画像データベースを想定した正解ラベル付き顔画像群を用いて，顔時間連続体の顔認証を行う（図 2 ⑥）．これ以降，本研究におけるマルチフレーム認識の枠組みを用いた顔認証の詳細を述べていく．

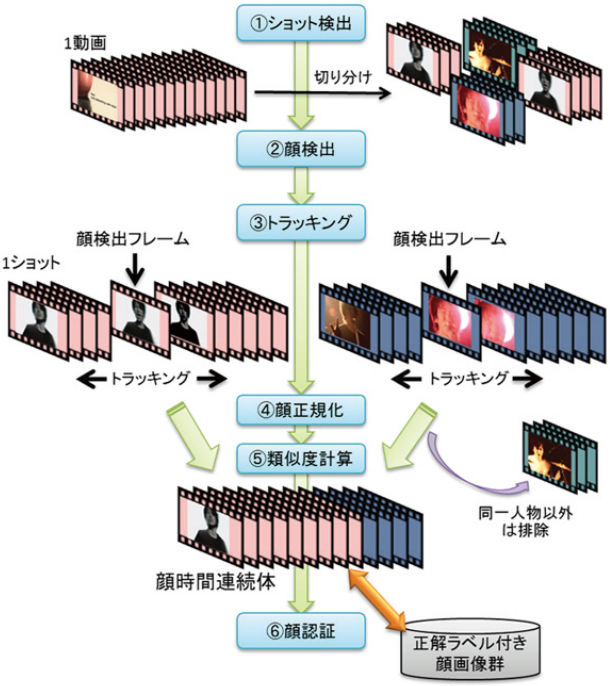


図 2　処理の概要
Outline of the method.

3.1　ショット検出

動画像において，シーンやカメラの切り替わりがなく，フレームが連続に繋がっている区間のことをショットという．動画像のフレーム連続性に基づけば，1 ショット中に映っている人物は，カメラや人物そのものの動きがなければ同一人物であると考えられる．また，カメラや人物の動きがあった場合にも，その動きを追いかけることで，同一人物をトラッキングすることができる．そのため，同一ショット内で人物の顔を検出した場合に，それらが時間的に連続していた場合，そのフレーム群は検出された同一人物の顔時間連続体として扱うことができる．そこで，まずは動画像のショットが切り替わる境界を自動的に検出し，得られた各ショットを顔時間連続体の候補とする．

本研究では，ショットの切り替わり箇所の判定のためにショット検出特徴量を用いる．総フレーム数 N の動画の各フレーム $i(i=1\sim N)$ に対して，画面輝度値 I のヒストグラム $H_i(I)$ を算出し，その 1 フレーム後のヒストグラムの値について，下の式 1 に示すショット検出特徴量 $D(H_i, H_{i+1})$ を元にショットの切り替わり箇所を判定する．

$$D(H_i, H_{i+1}) = \sum_I \frac{H_{i+1}(I) - H_i(I)}{H_{i+1}(I) + H_i(I)} \qquad (1)$$

このようにして算出したショット検出特徴量 $D(H_i, H_{i+1})$ は，**図 3** のようにショット切り替わり箇所以外では 0 に近い値をとり，動画像中の前後フレームにおいて画面の輝度値が大きく変化した箇所，すなわちショットが切り替わったフレームにおいてピークとして現れる．これに対し，ピークを検出するための閾値として，$\mu + 2 \times \sigma$ の値を設定した．ここで，μ，σ はそれぞれショット検出特徴量の平

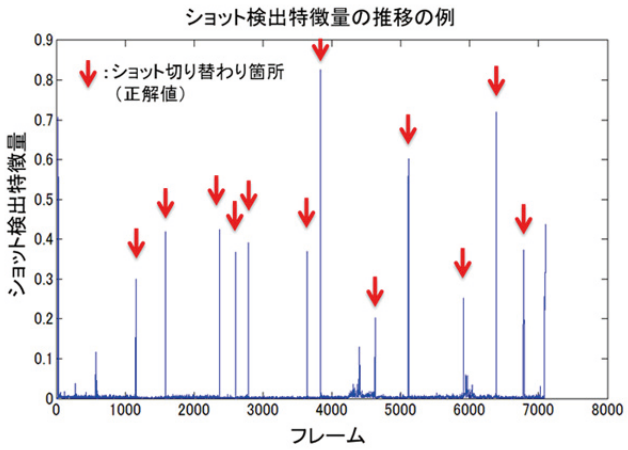


図 3　動画像中のショット検出特徴量の推移の例
An example of transition of shot detection feature in a video stream.

均値，標準偏差である．この閾値は本手法において重要なフレームの時間連続性を保障するために，検出漏れを抑え，過検出を許すような閾値設定となっている．

3.2　動画像中の顔検出

ショット検出により切り分けられた顔時間連続体の候補から，実際に顔が映っているフレームとその領域を検出する．動画像フレーム中の顔領域の検出には，Active Structure Appearance Model（ASAM）[6] による顔領域のグローバルフィッティングと，ローカルモデルによる顔部位毎のフィッティングを元に，顔検出を行う入江らの階層的フィッティング [7] の手法を用いた．

ASAM は，形状モデル上でのサンプリング点の構造的配置と特徴量による形状パラメータの摂動量を学習により関連づけることで，高速かつ高精度に顔輪郭点検出を実現する手法である．ASAM は，Active Appearance Model（AAM）や Active Shape Model（ASM）では困難であった学習されていない不特定多数の顔に対するフィッティングをリアルタイムで実現でき，動画コンテンツのような大量のフレームに対しての高速かつ高精度な顔検出を行うことができる．

しかし，ASAM は表情変化に対して誤検出を起こしやすいという欠点がある．そこで入江らは，階層的フィッティングを用いることにより，ASAM の欠点である表情変化に対してロバスト性を向上させている．これにより，表情や顔向きの変動が多い動画像においても高精度な顔検出を行うことができる．

この顔検出手法を，切り分けた動画像の全ショットに対して適用することで，各ショット中の人物顔と各顔器官の位置を検出し，それぞれのフレームで顔の特徴点 31 点を検出する．検出した顔特徴点の配置を**図 4** に示す．図 4 右から，表情変化に対応できていることがわかる．ただし，この階層的な ASAM によるフィッティングを用いても，顔の各部位のモデル形状を逸する極端な表情やオクルージョンには対応できない．その際には，顔領域は検出されるが，特徴点の誤検出を起こしてしまう．逆に，特徴点の検出精度さえ気にしなければ，オクルージョンや表情変化，その他の厳しい撮影条件に対してロバスト性の高い顔検出手法である．

図 4　検出した顔特徴点（緑）の配置
Example of face detection and detected facial feature points (green plots).

3.3　顔領域のトラッキング

各ショットで検出された顔領域の情報を元に，同一ショット内の顔が検出できなかったフレームに対しても顔領域がないか探索する．通常，ショットの切り替わり以外の箇所では，それまで映っていたはずの顔が次のフレームで消えるということは起こりづらい．そこで，顔検出が成功した前後のフレームには，同一の顔が映っている可能性が高いという仮定を置き，検出成功した顔領域を囲む正方形ブロックを探索ブロックとして，前後フレームに対してブロックマッチングを行う．ブロックマッチングの計算には，前後フレームの探索範囲の輝度値をそれぞれ I_1，I_2 として以下の式で表される Sum of Squared Difference(SSD）を用いた．

$$R_{SSD} = \sum_{i}^{width} \sum_{j}^{height} (I_1(i,j) - I_2(i,j))^2 \qquad (2)$$

ブロックマッチングを行った結果，R_{SSD} の値が最も小さい領域を顔の移動後の領域とする．ここで，探索結果の領域が顔かどうかを判定するために閾値を設定する必要がある．閾値の設定には，同一ショット内の前後 2 フレーム以上で顔検出が成功している連続フレームを用いた．この連続フレーム間の SSD の値を元に閾値を設定することで，ブロックマッチングによる顔のトラッキングを行った．同一ショット内の前後 2 フレームで顔が検出されていない単一顔フレームの情報を伝搬させる際には，他のショットにおけるトラッキング時に用いられた閾値の平均値を，ここでの閾値としてトラッキングを行う．このようにして，各顔検出の成功フレームの前後の方向に対して，顔情報を伝搬させながら探索していくことによって，顔検出が困難なフレームに対しても顔領域を検出していく．さらにここで，顔領域情報の伝搬と同様に顔特徴点の情報も伝搬させる．特徴点情報の伝搬は，顔検出が成功した顔フレームにおける顔領域の正方形ブロックと特徴点の間の位置関係を伝搬

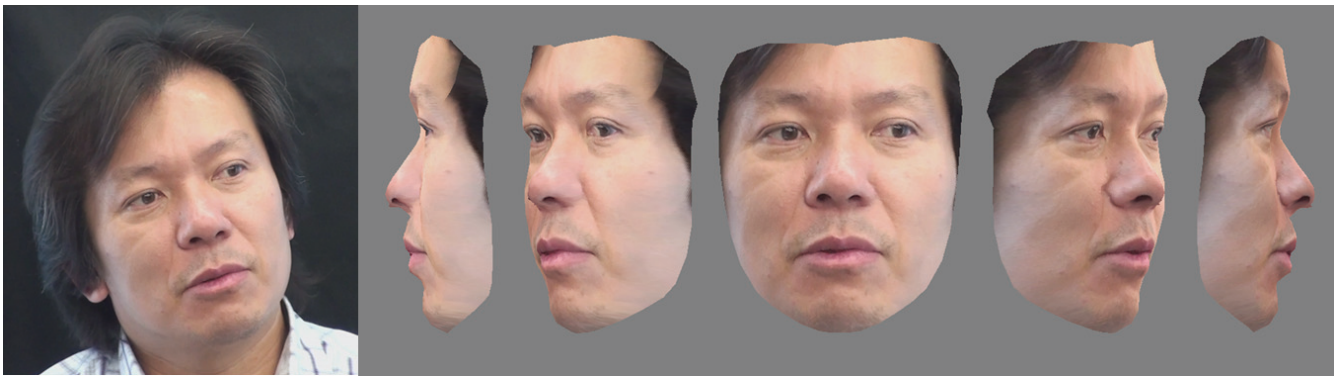


図 5　3 次元顔形状の復元による顔向き補正
Adjusting face direction by reconstructing 3D face form.

させることによって行う．本トラッキング手法では，ブロックのサイズを一定としたブロックマッチングを行っているため，カメラのズームなどが起こり，顔領域の大きさが変わる際にはトラッキングを続けることができない．

3.4　3 次元顔形状復元による顔向きの正規化

既存の顔認証手法では，撮影条件の違いを統一するために，正規化処理を加えたり，条件の違いに対して頑健な特徴量を用いたりすることで顔認証を行っていた．しかし，動画コンテンツにおける顔フレームの条件の違いには，照明，顔向き，表情などの通常の顔認証で直面する条件の他に，オクルージョン，経年変化，解像度の違いなどといった比較的困難な条件の違いもすべて同時に考慮しなければならない．

ここでは，動画コンテンツにおいて最も変動が大きい要素の一つである顔向きの正規化を行う．その他の顔認証の障害となる条件の違いについては，顔時間連続体の構築とクラスタリングによる様々な撮影条件の蓄積と，顔認証に使用する特徴量に撮影条件の違いへの頑健性を持たせることで対処する．ここで，顔向きのみを正規化の対象としたのは，顔向きは 3.2 節で述べた入江らの階層的フィッティングを用いた ASAM により，角度を算出することができ，その他の撮影条件に比べ，1 フレームのみの情報からでも正規化のための基準が得られやすいことによる．

顔向きの正規化は，顔向きに角度がある顔画像を正面顔に補正することで行う．画像の 2 次元平面における角度補正の場合，2 次元アフィン変換により角度の補正をすることができるが，顔が上下左右に傾いているような顔向きの補正を行うには，顔の 3 次元形状の復元を行う必要がある．そこで，2 次元顔フレームから 3 次元顔形状を復元するための手法として，Blanz らの統計的手法を用いた [8]．Blanz らの手法では，3 次元顔形状を学習データとし，2 次元テクスチャと 3 次元形状の間の特徴点の対応関係を学習することで，任意の 2 次元顔画像から 3 次元形状の復元を行っている．

Blanz らの手法を用いて，2 次元顔フレームから 3 次元形状を復元した例を**図 5** に示す．これにより，図 5 左に示した顔向きに傾きがある顔画像を，図 5 右に示した正面顔画像のように補正することができる．

3.5　顔領域の類似度判定とクラスタリング

ここまでに取得できた動画像中の顔領域と顔特徴点の情報を元に，顔の特徴を表す特徴量を算出し，特徴量間の距離を計算することで顔時間連続体間の類似度を算出する．複数のショットで構成される動画像において，同一人物の顔時間連続体は複数存在する可能性は高い．そこで，顔時間連続体の特徴量間類似度を測ることで，同一人物の顔時間連続体をクラスタリングして連結し，より長い顔時間連続体を得る．

ここで，類似度計算のために用いられるべき顔特徴量は，動画像における顔の変動と撮影条件に対する頑健性が求められる．ここまでの顔検出とトラッキング処理で得られた各顔フレームの特徴点の情報は，顔向きや表情変化に対してロバストであり，常に顔器官の同一の点を検出している．それを利用して，各特徴点の周辺の情報を特徴量として用いることで，顔時間連続体のクラスタリングを高精度に実現できると考えた．

そこで，我々は検出した顔フレームの各特徴点の周辺領域に対して Histogram of Oriented Gradient（HOG）を適用した [9)10)]．HOG 特徴量は輝度値の勾配情報を見るため，大域的な照明変動に対して頑健であり，多様な撮影条件に対する特徴量の変動を最小限に抑えることができるため，動画像に用いるのに適している．局所的な照明条件に対しても，特徴量の取得範囲を最小限に絞ることで，影響を受けづらくすることができる．具体的には，この HOG 特徴量を Scale-Invariant Feature Transform Invariant（SIFT）[11] のように，特徴点の周りのみに対して適用した．これにより，姿勢変動などにも頑健な同一箇所にのみ注目できる特徴量を構築する．HOG 特徴量は，複数のセルによって構成されるブロック領域において適用される特徴量であり，各セルの輝度勾配と輝度強度のヒストグラムを特徴量の値とするが，今回我々は各特徴点を中心とした 5 × 5 ピクセルをセル領域とし，検出された特徴点の周辺領域のみの HOG 特徴量を算出した．HOG 特徴量は一つのセルに対して 9

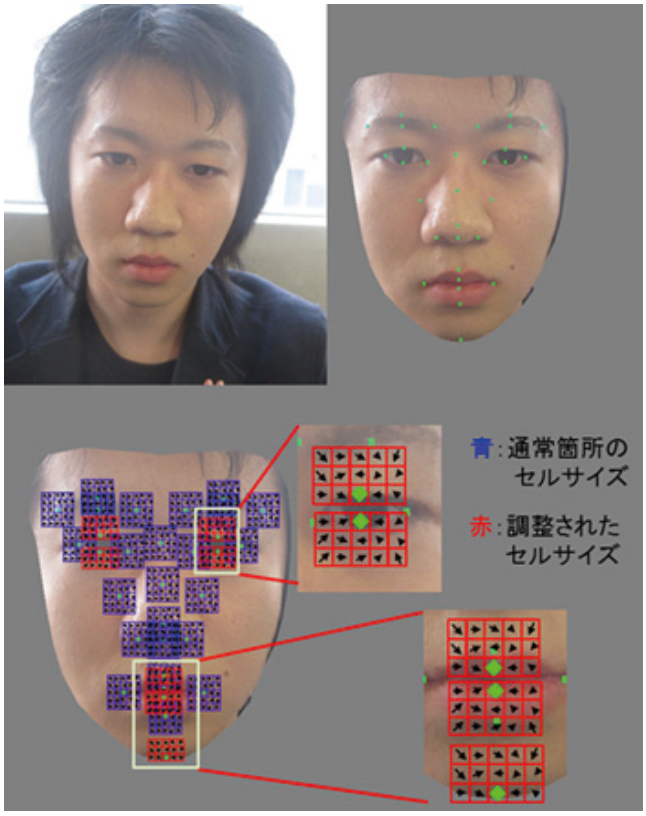


図 6　特徴点周辺の HOG 特徴量，および，セルサイズの調整による表情変化への対応
HOG feature around the facial feature points and reactions to facial expression by changing cell size.

個のビンを持つヒストグラムにより記述されるため，1 つの特徴点に対して 9 次元の特徴量として表される．これを，31 特徴点を中心としたすべてのセルに対して算出し，279 次元の特徴量として記述する．特徴点周辺の HOG 特徴量抽出の様子を**図 6** に示す．ここで，まぶたの上下の特徴点や，上唇の下側，下唇の上側の特徴点は，目や口の開閉による輝度勾配の変化が起こる．そこで，表情変化に頑健な特徴量とするため，表情変化による特徴点周辺情報の変化が起こりやすい特徴点においては，図 6 右のようにセルの大きさを上領域または下領域のみの半分の大きさにして特徴量の抽出を行った．これらの処理により，照明条件と表情変化に対してロバスト性を持った顔の特徴量を抽出している．

このようにして算出した特徴量の間の類似度を測る．類似度は，各顔フレームにおける特徴量 279 次元間のユークリッド距離を計算することによって算出する．これを全顔検出フレーム間で計算する．ここで閾値を設定し，閾値以上の類似度を示した顔フレームが n フレーム以上ある場合には，それらの顔時間連続体同士は同一人物であると判定する（現在の実装では，予備実験の結果 $n = 5$ とした）．また，同一人物間であっても，類似度が高くない限りは同一人物としないように閾値を設定している．これによって，異なる人物をクラスタリングすることを防いでいる．

この類似度計算を動画像全体および，複数の動画間で行うことで，同一人物の顔時間連続体をクラスタリングすることができる．クラスタリングができなかった顔時間連続体は単一の顔時間連続体として保持する．この単一顔時間連続体は，動画像中の 1 シーンのみに出現した可能性のある人物や，クラスタリングされた顔時間連続体とは例外的に異なる条件で撮影された同一人物の可能性がある顔時間連続体である．後者の場合，同一人物に関する動画像サンプルに対してさらに分析していくことで，顔時間連続体同士が結合される．

本手法は，フレームの時間連続性を利用したマルチフレームを蓄積することによって，様々な条件下での顔情報を収集するため，例えば，オクルージョンなどが存在するフレームがあったとしても，それに続くフレームでオクルージョンがなくなっている場合に，オクルージョンの影響を受けなかったフレーム同士で高い類似度を示すフレームが存在しうるという特性がある．これにより，時間的に連続しているフレームのどこかに，より認証に適した条件のフレームが存在すれば，様々な条件も吸収することができる手法となっている．

3.6　顔認証

ここまでに得られた同一人物に関しての膨大な顔時間連続体と，クラスタリングされなかった単一顔時間連続体群に対して顔認証を行う．これらの顔時間連続体は同一人物の様々な顔向き，表情を集めたデータとなっており，特にクラスタリングされた顔時間連続体については，様々な撮影条件も吸収した顔データとなっている．そのため，これらの連続体を用いて通常の顔認証と同様のステップとして，特徴量間の類似度計算を行えば，少ない正解顔画像からでも，あらゆる条件に対してロバストな顔認証を行うことができる可能性がある．ここで，認証のために使用した特徴量は，比較を行う 2 枚の顔画像において，正規化した顔領域の両目間の中点を中心として，ちょうど顔領域がおさまるサイズの長方形ブロックの画素値の差である．正解データと各顔フレームの類似度は，特徴量のユークリッド距離によって算出する．この顔認証のステップにより，膨大な数の同一人物の顔フレームすべてに対して，一度に正解ラベルを付与することができる．顔認証のために新たな特徴量を使用したのは，膨大なフレーム数をもつ顔時間連続体を構築した後では，顔時間連続体にその人物の有名なシーンが含まれる可能性も考えられ，インターネットから取得する正解顔画像の中に，まったく同一の瞬間の顔画像が含まれる場合が考えられるからである．

認証対象とする正解顔画像は，Faces in the wild と呼ばれる様々な人種や年齢の人物の一般条件下で撮影された顔画像を集めた正解ラベル付きのデータセットの中から，ランダムに 500 枚を選んだ [12]．この 500 枚の中に，対象動画コンテンツに映っている人物の正解顔画像も含めた．この正解顔画像群に対して，顔時間連続体の全フレームを入力クエリとして，各フレームに対して最も類似度の高い顔

表 1　顔検出（3.2 節）の性能とトラッキング（3.3 節）の効果
Performance of face detection and effect of tracking.

動画名	人物総出演フレーム数	顔検出の性能（トラッキング前）			顔検出の性能（トラッキング後）		
		顔検出成功フレーム数	顔誤検出フレーム数	顔検出率 [%]	顔検出成功フレーム数	顔誤検出フレーム数	顔検出率 [%]
Let it be / The Beatles	7073	3524	0	100.00	3610	1	99.97
Hey Jude / The Beatles	7119	3678	1	99.97	3871	1	99.97
Get Back / The Beatles	4957	848	1	99.88	935	1	99.89
Two of us / The Beatles	6136	1628	0	100.00	1717	0	100.00
The Beatles の全動画	25285	9678	2	99.98	10102	3	99.97
Can You Keep A Secret?/宇多田ヒカル	4208	1284	21	98.39	1446	39	97.37
Wait & See / 宇多田ヒカル	3974	2271	72	96.93	2404	98	96.08
For You / 宇多田ヒカル	5447	1309	7	99.47	1486	11	99.27
Final Distance / 宇多田ヒカル	3905	489	3	99.39	553	6	98.93
宇多田ヒカルの全動画	17527	5353	103	98.11	5889	154	97.45

画像を探索する．さらにそれらのフレームと顔画像の組合せの中で，最も類似度の高い顔画像を顔時間連続体の正解顔画像とすることで，顔時間連続体全体に対して正解ラベルを付与する．

4. 実験と結果

4.1 実験条件

本手法を用いて，実際の動画像に対して同一人物のクラスタリングおよび，顔認証を行った．実験にはライヴ映像やPV（Promotion Video）などの音楽動画コンテンツ（本研究では，音楽連動動画と呼ぶ）を使用した．音楽連動動画は，作品毎に演出やメイクが変わり，動画コンテンツの中でもシーンの切り替えが非常に多い．そのため，音楽連動動画では，各ショットの撮影条件の違いや顔向きの変動が大きく，通常の画像同士を比較する顔認証では，各ショットに対して適切なモデルの構築を行う必要があり，非常に困難な題材である．演奏動画では，シーンが変化することはほとんどないが，ショットの切り替わりが多い．そのため，照明の条件は安定している．一方 Promotion Video ではショットの切り替わりだけでなく，シーンの切り替わりも多く，照明の変化も大きい．本手法は，人物が映っている動画コンテンツ全般に対して適用できるが，撮影条件の分散が大きい題材として，音楽連動動画を用いて実験を行った．このような分散の大きい動画コンテンツに対して本手法の有効性を示すことができれば，人物顔が映っている一般的な動画コンテンツ全般に対しても本手法が有効であると考えられる．

今回実験に用いたのは，The Beatles による演奏動画 4 作品，宇多田ヒカルの Promotion Video4 作品の計 8 作品である．ここで，The Beatles の 4 動画は，撮影された年代にばらつきを持たせ，人物に経年変化が起こっているサンプルを用いた．実験手順として，まず各動画中の同一人物をクラスタリングした後，アーティスト（人物）毎に類似度計算を行い，顔時間連続体を構築する．その後構築した各アーティストの顔時間連続体に対して顔認証を行う．ここで使用するのは，Face in the wild データセットよりランダムに取得した 495 枚の顔画像に，The Beatles のメンバー 4 人と宇多田ヒカルの正解顔画像 1 枚ずつを含めた計 500 枚の名前ラベル付きデータベースである．

4.2 実験結果と考察

実験結果のうち，3.2 節で記述した階層的フィッティングを用いた ASAM による顔検出の結果と 3.3 節で記述した顔領域のトラッキングの結果を**表 1** に示す．トラッキングの前後で顔検出フレーム数が増加しているが，顔検出率はほとんど変化していないことがわかる．このことから，フレームの時間連続性を利用したトラッキングにより認証対象となる顔フレームを，精度を維持したまま増加させることができたといえる（検出フレーム数の上限は，人物総出演フレーム数である.）また，**表 2** に顔認証結果を示す．ここで，認証率は以下の式で表されるものとする．

$$\text{認証率} = \frac{\text{認証成功フレーム数}}{\text{トラッキング後の顔検出フレーム数}} \quad (3)$$

認証率は，本手法により認証を行う対象となったフレーム（トラッキング後の顔検出フレーム）に対する正解ラベルが付与された認証成功フレームの割合である．トラッキング後の顔検出フレームには，表 1 に示したような誤検出フレームも含まれている．比較のために，各顔検出フレームに対して 500 枚の顔画像群の中から顔認証を行った結果も示している．この結果から，The Beatles の全動画（演奏動画）においては約 2 倍，宇多田ヒカルの全動画（Promotion Video）においては約 3 倍の認証率の向上を実現したことがわかる．

さらに，**表 3** には，全顔連続体における認証成功した連続体の割合を示す．この結果から，The Beatles の全動画では，認証成功した連続体数が宇多田ヒカルの全動画に比べて少ないことがわかる．これは，経年変化の影響で顔時間連続体同士のクラスタリングが多くは行われなかったことが原因であると考えられる．しかし，一番長いフレームにおいて登場していたボーカルに対しては，3 段階の経年変化顔に対しても適切なクラスタリングが実現された．これは，本手法では顔の類似度を局所的な HOG 特徴量を元に算出しており，特徴点の周辺の局所的な特徴が経年変化を起こしにくいことに起因するのではないかと考えられる．経年変化への適用可能性については，別途検討していく余

表 2 顔認証率の比較 (フレーム数での比較)
Comparison of face recognition ratios (comparison of the number of frames).

動画	トラッキング後の顔検出フレーム数	単一フレーム毎の認証成功フレーム数	顔時間連続体による認証成功フレーム数	単一フレーム毎の認証率 [%]	顔時間連続体の認証率 [%]
The Beatles 全動画	10136	2871	6489	28.3	64.0
宇多田ヒカル全動画	6043	1960	5778	32.4	95.6

表 3 顔時間連続体単位での認証率
Face recognition ratio with facial-temporal continuum.

動画	顔時間連続体数	認証成功連続体数	連続体単位での認証率 [%]
The Beatles の全動画	157	30	19.1
宇多田ヒカルの全動画	295	271	82.0

表 4 顔時間連続体におけるフレーム単位でのクラスタ誤り率
Average of error rate in each facial-temporal continuum.

動画	トラッキング後の顔検出フレーム数	クラスタ誤りフレーム数	クラスタ誤り率 [%]
The Beatles の全動画	10136	46	0.5
宇多田ヒカルの全動画	6043	149	2.5

地がある．さらに，一番長いフレームに登場していた人物に関しては，マイクなどのオクルージョンや，髭の有無などの条件に対しても適切なクラスタリングが行われた．

また，認証成功した連続体が持つ平均フレーム数は 1 連続体当たり 45.1 フレームで，認証失敗した連続体の平均フレーム数は 20.5 フレームであり，認証成功した連続体の方が多くのフレームを含有している傾向があった．このことから，顔認証をする上で，多くのフレーム情報を持っていた方が認証の精度は高くなるということがいえる．

表 4 にはフレーム単位で見た際の，顔時間連続体において誤ってクラスタリングされたフレーム数とその割合を示す．この結果から，全体のフレーム数では，顔時間連続体はおおむね 97 %以上の割合で同一人物をクラスタリングできているということがわかる．顔時間連続体の構築によって誤った人物を同一人物としてしまう誤り率に対して，顔認証率の向上の方が大きく，本研究のフレーム連続性と顔類似度を用いた枠組みによって，動画像における顔認証の精度が向上していることがわかる．

表 5 にはクラスタ単位で見た，顔時間連続体同士のクラスタリングの個数と誤ってクラスタリングされた割合を示す．ここで，クラスタ誤りを起こしている連続体に注目すると，長時間にわたって特徴点の誤検出が起こってしまっている．階層的フィッティングを用いた ASAM による顔検出で，顔の表情変化や傾き，オクルージョンへのロバスト性は高くなっているが，極端な表情変化，極端なオクルージョンが存在するフレームでは，顔検出が成功しても顔特徴点の誤検出が起こってしまう．本手法において，顔の類似度は特徴点の周辺情報を元に算出しているため，継続した特徴点の誤検出により，クラスタ誤りが起こってしまう．

本手法により，認証精度が向上した箇所，誤認証の原因となった箇所が存在する．認証精度が向上した箇所は，通

表 5 顔時間連続体間のクラスタ誤り率
Average of error rate between facial-temporal continua.

動画	クラスタリングされた連続体数	クラスタ誤り連続体数	クラスタ誤り率 [%]
The Beatles の全動画	37	8	21.6
宇多田ヒカルの全動画	160	5	3.1

常の顔検出では顔を検出できなかったフレームのうち，フレームの時間連続性を利用して顔トラッキングを行うことができたフレームである．本手法のフレームの時間連続性による顔時間連続体の構築により，単一フレーム毎には顔を検出することさえできなかったフレームについても認証を行うことに成功した．また，本手法では，顔時間連続体を構成するマルチフレームのうち，最も撮影条件の良い顔フレームを認証に使用して，撮影条件の悪い顔フレームにまで正解ラベルを付与することができるため，表 2 に示すように認証精度が向上している．逆に本手法により，誤認証の原因となった箇所としては，クラスタ誤りにより，異なる人物であるにも関わらず，同一人物としてクラスタリングされてしまった顔時間連続体のフレームである．この点は，単一フレームでの顔認証ならば，それぞれのフレーム毎に顔認証をできるため，正解ラベルを付与できる余地が残っている．この問題に関しては，顔時間連続体のクラスタリング精度を向上させることで解決できる可能性がある．

5. むすび

本研究では，動画コンテンツにおけるマルチフレーム認識の枠組みを用いることで，顔に特化した認証手法を提案し，その有効性を確かめた．多くの動画コンテンツにおいて人物顔は動画像の主となる重要な情報である．本手法により認証した顔情報を元に，分析範囲を人物の体全体や周辺情報にまで広げていくことで，動画像の内容理解をより詳細に行うことができると考えられる．一般的に，後ろ向きの人物に対する顔認証は困難な課題であるが，提案手法を応用していくことで実現できる可能性があり，それを目指したい．

本手法のような，マルチフレームを蓄積していく手法では，データの蓄積が多くなるとデータの分散も大きくなり，認証やクラスタリングの誤りに繋がってしまう．本手法では，顔の類似度を算出する上での特徴量を，局所的な 279 次元の情報とし，ある程度のデータの分散に対してもクラスタリングの誤りが起こりづらくなるようにしている．さらに，顔時間連続体同士をクラスタリングする際の閾値を

手動で実験的に決定しており，それによりデータの分散を抑制している．しかし，扱うデータの量が膨大になっていくと，手動で決定した閾値では不充分となりうる．そこで，今後はクラスタ誤りを抑えるような閾値自動決定手法の導入が必要であると考えている．

現在は，顔時間連続体の正解顔画像を含む正解顔画像群を手動で用意した後に顔認証を行っているが，これは，インターネット上の膨大な顔画像群を活用することを想定している．今後，さらに認証精度を向上させることで，インターネット上に無数に存在する顔画像を利用して，手動で正解ラベルが付けられないような人物に対しても正解ラベルを付与することを考えている．

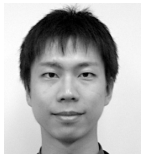
また，顔認証の際にソーシャルアノテーション情報を利用することで，さらなる精度向上が可能であると考えている．ソーシャルアノテーションから動画コンテンツに映っている人物の候補を導くことができるほか，出演者の出現確率などを利用して，顔認証の精度をさらに向上させることができると考えている．

今後，ソーシャルアノテーションの利用による認証精度の向上を図るとともに，この手法の応用範囲を拡張していくことで，顔以外のオブジェクトの認識やそれを元にした映像理解をするための手法についても検討していく．また，出演の予測がしがたい人物などに対する顔認証も図れるようなユーザによる補正を加えられる枠組みなど，インタラクション性を持つシステムに関する研究も進めていきたい．

〔文　　献〕

1) 佃洸摂，中村聡史，山本岳洋，田中克己：“映像に付与されたコメントを用いた登場人物が注目されるシーンの推定”，情処論，**52**，12，pp.3471-3482 (2011)
2) TRECVID : http://www-nlpir.nist.gov/projects/trecvid/
3) 樋爪和也，柳井啓司：“マルチフレーム認識を用いた動画像認識の分析”，情処学研究報告，CVIM，**177**，28，pp.1-8 (2011)
4) 滝沢圭，長谷部光威，助川寛，佐藤俊雄，榎本暢芳，入江文平，岡崎彰夫：“歩行者顔照合システム「FacePassenger^{TM}」の開発”，情報科学技術フォーラム一般講演論文集，**4**，3，pp.27-28 (2011)
5) 山名信弘，井辺昭人，三浦文裕，前島謙宣，森島繁生：“動画の 3 次元周波数成分を用いた顔認証システム”，信学技報，PRMU，**106**，73，pp.13-18 (2006)
6) 木下航一，小西嘉典，勞世　，川出雅人，村瀬洋：“摂動特徴量による顔画像に対する形状モデルフィッティング”，信学論，**J94-D**，4，pp.721-729 (2011)
7) A. Irie, M. Takagiwa, K. Moriyama, T. Yamashita :“Improvements to Facial Contour Detection by Hierarchical Fitting and Regression”, Asian Conference on Pattern Recognition (ACPR), pp.273-277 (2011)
8) V. Blanz, A. Mehl, T. Vetter, H. Seidel:“A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data”, Symp. on 3D Data Processing, Visualization, and Transmission, pp.293-300 (2004)
9) N. Dalal, B. Triggs :“Histograms of Oriented gradients for human detection”, Proc. IEEE, Conference on Computer Vision and Pattern Recognition (CVPR), pp.886-893 (2005)
10) 大戸和博，土肥慶亮，柴田裕一郎，小栗清：“HOG 特徴と AdaBoost による人検出処理の FPGA への実装”，信学技報，CPSY，**110**，360，pp.117-122 (2011)
11) D. G. Lowe:“Object Recognition from Local Scale-Invariant Features”, Proc. IEEE, International Conference on Computer vision (ICCV), pp.1150-1157 (1999)
12) T. L. Berg, A. C. Berg, J. Edwards, D. A. Forsyth:“Who's in the Picture”, Proc. Neural Information Processing Systems (NIPS), pp.137-144 (2004)

平井　辰典（ひらい　たつのり） 2011 年，早稲田大学先進理工学部卒業，理学学士，現在に至る．情報処理学会音楽情報科学研究会会員．2010 年より，音楽情報処理，動画像処理に関する研究に興味を持つ．



中野　倫靖（なかの　ともやす） 2008 年，筑波大学大学院図書館情報メディア研究科博士後期課程修了．博士（情報学）．同年，産業技術総合研究所に特別研究員として所属し，現在，産業技術総合研究所研究員．情報処理学会，日本音響学会各会員．平成 21 年度 情報処理学会 山下記念研究賞等，3 件受賞．

後藤　真孝（ごとう　まさたか） 1998 年，早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年，電子技術総合研究所に入所し，2001 年に改組された産業技術総合研究所において，現在，情報技術研究部門 上席研究員 兼 メディアインタラクション研究グループ長．統計数理研究所 客員教授，筑波大学大学院 准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業プロジェクトマネージャーを兼任．ドコモ・モバイル・サイエンス賞 基礎科学部門 優秀賞，科学技術分野の文部科学大臣表彰 若手科学者賞，情報処理学会 長尾真記念特別賞等，29 件受賞．

森島　繁生（もりしま　しげお） 1987 年，東京大学大学院工学系研究科博士課程修了，博士（工学）．同年，成蹊大学工学部専任講師．1988 年，同助教授．2001 年，同電気電子工学科教授．2004 年，早稲田大学理工学部応用物理学科教授．早稲田大学 IT 研究機構セキュリティ・セイフティ研究所所長．現在に至る．明治大学理工学部，新潟大学非常勤講師を併任．1991 年，電子情報通信学会業績賞，2010 年，電気通信財団テレコムシステム技術賞受賞．画像電子学会理事，日本顔学会理事．正会員．