Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

音楽動画コンテンツ中のアーティスト名とその登場シーンの同定手法

平井辰典† 中野倫靖†† 後藤真孝†† 森島繁生†

本稿では，音楽動画コンテンツに対して「どのアーティストがいつ映像中に登場しているか」というアノテーション情報を自動付加する手法を提案する．従来の人物顔認証手法は映像中の照明や顔向きなどの撮影環境の変動に脆弱で，その変動が大きい音楽動画コンテンツにおいて，アーティスト名とその登場シーンを同定することは困難であった．そこで本研究では，映像のフレームの時間的連続性を利用して同一人物の顔をクラスタリングすることで，撮影環境の違いを吸収し，アーティストの顔認証をおける問題を解決した．本手法により，従来の単一フレーム毎に顔認証を行う手法に比べ，約 2〜3 倍の精度向上を実現した．また，音楽の歌声区間と映像中にボーカリストが登場するシーンとの関係についても調査し，それを利用した今後の精度向上の可能性について考察した．

A Method to Identify Artist's Name and Its Performance Scenes in Music Video Content

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

In this paper, we propose a method that can automatically annotate when and which artist is appearing in a music video clip. Previous face recognition methods were not robust against different shooting conditions such as variable lighting and face directions in a music video clip, and had difficulties identifying artist's name and its performance scenes. To overcome such difficulties, our method groups consecutive video frames (scenes) into clusters each having the same artist's face, and identifies an artist 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 recognizing a face in each frame. Furthermore, we discuss possible improvements by using relationship between the appearance of a vocalist in a video clip and sung sections in its song.

1. はじめに

近年，インターネット上で音楽に関連した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いる．本研究では，そうした音楽動画コンテンツの中でも音楽内容に密接に連動した動画（アーティストのミュージックビデオ，ダンス動画等）を「音楽連動動画」と呼ぶ. 例えば 2012 年 1 月の時点で, 動画検索サービス「Google Videos[1)]」において「music」というキーワードで検索すると約 13 億 8 千万動画，「音楽」で検索すると約 570 万動画がヒットする．動画コミュニケーションサービス「ニコニコ動画 [2)]」でも，総動画数の約 23%が音楽に関連した動画である．その多さから，ユーザが興味を持つ音楽動画コンテンツを的確に探して閲覧することが困難である．

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．音楽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．テキスト情報でもアーティスト名や曲名などの「誰が映っている何についての映像か」は判明することが多いが，音楽連動動画では時系列情報，例えば，「いつ誰が何をしているのか」という情報も重要となる．音楽連動動画に時系列のアノテーション情報を付加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加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．

本稿では，音楽連動動画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アーティスト名（本稿では，アーティストがグループの場合には，その個人名を意味する）とその登場シーンに焦点を絞り，映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加する新しい手法を提案する．そして本手法の実装方法と，アーティスト名とその登場シーンの同定実験を行った結果を述べ，その有用性を議論する．

本手法により，ユーザは好きなアーティストが映っている音楽連動動画を横断的に検索し，視聴することが可能となる．また，アーティストの 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音楽連動動画の中からそのアーティストが写っているシーンを検索して抜き出すといった，より詳細な動画検索が可能となる．

† 早稲田大学先進理工学研究科
Graduate School of Advanced Science and Engineering, Waseda University
†† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

2. 研究背景

音楽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，映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，佃らが，ニコニコ動画の視聴者によって付与された時刻に同期したコメント情報に基づいて，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 [3]．しかし，ソーシャルアノテーションには主観的な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，映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．

そうした映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 [4]（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・分類した結果，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 [5]. TRECVID では物体認識を元に映像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．本研究が対象とする音楽連動動画に対する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（アーティスト）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，映像中の人物の顔を認識（認証）する研究も多く行われている [6),7)]．しかし撮影環境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環境を限っていることが多かった．そのため，音楽連動動画のように，アーティストの顔が映っているフレーム（以降，顔フレームと呼ぶ）が多様な撮影環境下で登場する用途には利用できない．例えば，従来は同一撮影環境下の映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環境下で顔認証をするには，それぞれの撮影環境に合わせたモデルの構築が必要で現実的でない．

既存の顔認証手法でも，補正や正規化の処理を加えたり，撮影環境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，音楽連動動画の顔フレームでは，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ず，難しい．さらに，音楽連動動画では，作品毎に演出やメイクが変わり，シーンの切り替えが多い．シーン毎に撮影環境の違いや顔向きの変動も大きく，単に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．

そこで本研究では映像中のフレームの時間的な連続性に着目し，同一人物の様々な顔フレームを関連づけて蓄積することで，多様な撮影環境のシーンを含む音楽連動動画対しても有効に機能する人物顔のマルチフレーム認識の手法を考案した．さらにその認識過程で生成される人物顔のマルチフレーム情報を活用して，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．

3. シーンの連続性と顔類似度に基づく音楽連動動画中の同一アーティストの同定手法

本研究では，音楽連動動画を対象として，まず映像中のアーティストの顔をトラッキングして顔フレームを自動推定し，続いて顔認証によってそのアーティスト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アーティスト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．本稿では，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影環境の違いを吸収して,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．

音楽連動動画に限らず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，全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．特に音楽では，ライヴ映像では撮影環境が悪い場合が多い，アーティストのミュージックビデオではそのビデオでしか使われないようなコスチュームや見た目（眼鏡の着用，等）であることがある等，条件はさらに悪い．そのため，映像の全フレームについて，個々のフレームだけに着目して，顔領域を検出することは困難である．

それに対して本手法では，映像のフレームの時間的な連続性に注目することで映像中の顔領域を検出する．フレームの時間的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，映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，映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うことができれば，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認証し

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

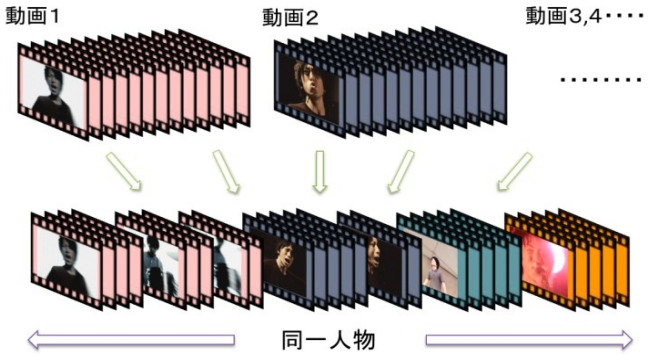

図 1　顔時間連続体の概念図

たい本人の名前（以降，正解ラベルと呼ぶ）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）によって音楽連動動画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及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．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．

3.1　ショット検出

映像において，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．映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，同一ショット内でアーティスト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，まずは映像をショットが切り替わる境界を自動的に検出し，そのようにして得られた各ショットを顔時間連続体の候補とする．

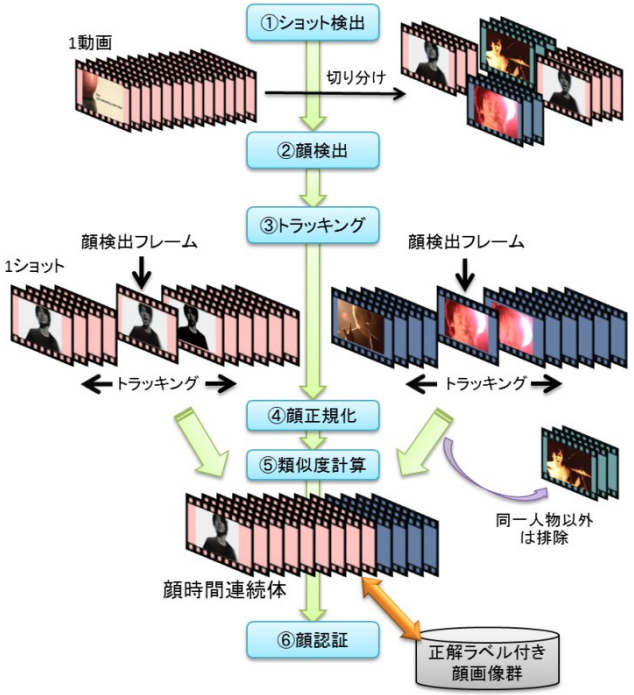

図 2　処理の概要

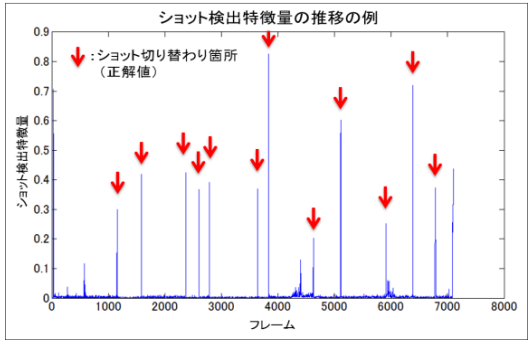

図 3　映像中のショット検出特徴量の推移の例

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

本稿では，ショットの切り替わり箇所の判定のために，従来よく知られた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，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．

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

このようにして算出したショット特徴量$D(H_i, H_{i+1})$は，図 3 のように映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．これを閾値処理により判定する．

3.2 映像中の顔検出

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．映像フレーム中の顔領域の検出には，Active Structure Appearance Model(ASAM)[8]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[9]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コンテンツ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，切り分けた映像の全ショットに対して適用することで，各ショット中のアーティストの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．

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(2)で表される Sum of Squared Difference(SSD)を用いた．

図 4　検出した顔特徴点（緑）の配置

$$R_{SSD} = \sum_i^{width} \sum_j^{height} (I_1(i,j) - I_2(i,j))^2 \quad (2)$$

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させることによって行う．

3.4 3 次元顔形状復元による顔フレームの正規化

既存の顔認証手法では，撮影環境の違いを統一するために，正規化処理を加えたり，環境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，音楽連動動画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

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

った比較的困難な条件の違いをもすべて同時に考慮しなければならない．

ここでは，音楽連動動画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環境におけるデータの蓄積と，顔認証に使用する特徴量に撮影環境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により，角度を算出することができ，その他の撮影環境に比べて，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 次元平面内での角度補正の場合，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 [10]．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映像中の顔領域と顔特徴点の情報を元に，顔の特徴量を算出し，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ている映像において，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には映像における顔の変動と撮影環境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，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[11),12)] (HOG)を適用した．HOG 特徴量は照明変動に対して頑健で，多様な撮影環境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，動画コンテンツに用いるのに適している．この HOG 特徴量を，Scale-Invariant Feature Transform [13)] (SIFT) のように，特徴点の周りのみに対して適用することで，姿勢変動などにも頑健な同一箇所のみに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

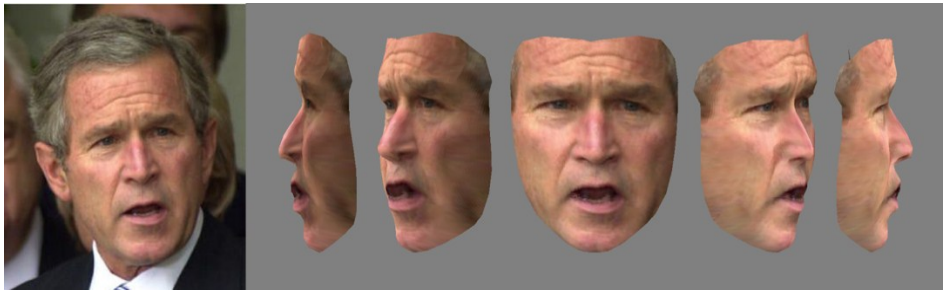

図 5　顔向き画像の 3 次元顔形状復元

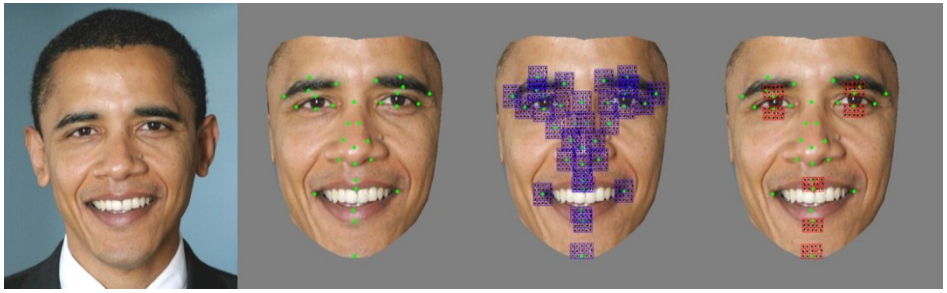

図 6　特徴点周辺の HOG 特徴量とセルサイズの調整による表情変化への対応

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 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，計 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として，全顔検出フレーム間で計算する．その際閾値を設定し，閾値以上であった顔フレームが 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してクラスタリングする．現在の

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

実装では，予備実験の結果 n=5 とした．また，異なるアーティスト同士をクラスタリングすることを防ぐため，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実験的に閾値を設定した．

この類似度計算を映像全体，及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，映像中の 1 シーンにのみ出現したアーティストや，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環境で撮影された同一アーティストである可能性がある顔時間連続体である．後者の場合，同一のアーティストに関する映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．

3.6 顔認証

ここまでに得られたクラスタリングによる同一人物に関する膨大な顔時間連続体のまとまりと，クラスタリングされなかった単一顔時間連続体群のすべてに対してアーティスト名同定のための顔認証を行う．これらの顔時間連続体は同一人物の様々な顔向き，表情によって構成され，特にクラスタリングによって連結された顔時間連続体は，様々な撮影環境を含む顔データといえる．そのため，これらの連続体を用いて通常の顔認証と同様に特徴量間の類似度計算を行えば，環境の違いに対してロバストな顔認証を行うことができる．ここで，認証のために使用した特徴量は，正規化した顔領域の両目間の中点を中心として，顔領域がおさまるサイズの長方形ブロックの画素値である．認証したいアーティストの正解顔画像と各顔フレームの類似度は，こ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，正解ラベルを一度に付与することができる．この顔認証のステップは，顔時間連続体を検索クエリとして，インターネット上に存在する無数の正解ラベル付き顔画像を対象に行うことを想定している．有名なアーティストの場合，膨大なフレーム数をもつ顔時間連続体の中にそのアーティストの有名なシーンが含まれる可能性も考えられる．顔認証のために新たな特徴量を使用したのは，インターネットから取得した正解顔画像候補の中にまったく同一の瞬間の顔画像が含まれる場合も考えられるからである．

本手法では，インターネットから取得する正解顔画像候補と顔時間連続体を比較することで，顔認証を行うことを想定しているが，今回は便宜上，あらかじめ正解顔画像候補となる顔画像群を用意して実験した．正解顔画像候補には，Faces in the wild と呼ばれる様々な人種や年齢の人物の一般環境下で撮影された顔画像を集めた人名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 [14]．この 500 枚の中に，対象の音楽連動動画に映っているアーティスト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画像を探索する．さらにそれらのフレームと顔画像の組み合わ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

本手法を用いて実際の映像に対して同一人物のクラスタリング及び，顔認証を行った．実験には音楽連動動画の中でも，アーティスト本人が出演している演奏動画（ライヴ動画等）やミュージックビデオ（Promotion Video を略して PV と呼ばれることもある）を使用した．今回実験に用いたのは，The Beatles による演奏動画 4 作品，宇多田ヒカルのミュージックビデオ 4 作品の計 8 作品である．まず，各動画中の同一人物をクラスタリングした後，アーティスト毎に類似度計算を行い，顔時間連続体を構築する．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，Faces 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 枚の人名ラベル付きデータベースである．

実験結果を表 1〜2 に示す．まず，全動画の全フレームに対して，各フレームの登場アーティストを人手によりラベリングした．それにより全人物の総出演フレーム数を算出した. これは, 人物が映っていると認識できるすべてのフレームのことであり，後ろを向いている人物のフレームや一部分だけが映っている人物のフレームも含まれている．表 1 の顔検出フレーム数は 3.2 節で行った顔検出の結果であり，トラッキング後の顔検出フレーム数は 3.3 節におけるトラッキング処理を行った後の顔フレーム検出数を表している．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ことがわかる．表 2 には顔認証率を示す．ここで，認証率は以下の式(3)で表すものとする．

$$\text{認証率} = \frac{\text{正解フレーム数}}{\text{トラッキング後の顔検出フレーム数}} \tag{3}$$

比較のために，各顔検出フレームにそれぞれに対して 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，単一フレームでは認証が困難な顔フレームに対しても，映像の時間的連続性を用いた顔時間連続体として扱うと認証精度は大幅に向上するということがいえる．

さらに，表 3 に，顔時間連続体のクラスタリングの誤り率を示す．これは，各顔時間連続体において，異なる人物を同一人物としてクラスタリングしてしまった割合であり，以下の式(4)で表す．

$$\text{誤り率} = \frac{\text{誤クラスタリングフレーム数}}{\text{トラッキング後の顔検出フレーム数}} \tag{4}$$

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

表 1　顔認証実験結果

動画名	総フレーム数	人物総出演フレーム数	顔検出フレーム数	トラッキング後の顔検出フレーム数
Let it be / The Beatles	7097	7073	3524	3611
Hey Jude / The Beatles	7196	7119	3681	3871
Get Back / The Beatles	5459	4957	849	936
Two of us / The Beatles	6218	6136	1628	1687
The Beatles の全 4 動画	25970	25285	9682	10105
Wait & See / 宇多田ヒカル	6655	3974	2343	2504
For You / 宇多田ヒカル	6034	5447	1316	1476
Can You Keep A Secret? / 宇多田ヒカル	5931	4208	1305	1485
Final Distance / 宇多田ヒカル	4924	3905	492	559
宇多田ヒカルの全 4 動画	22570	17527	5456	6024

表 2　顔認証率

動画名	トラッキング後の顔検出フレーム数	正解フレーム数	顔時間連続体の認証率[%]	フレーム毎の認証率[%]
The Beatles の全 4 動画	10105	6489	64.2	28.4
宇多田ヒカルの全 4 動画	6024	5778	96.0	32.5

表 3　顔時間連続体内でのクラスタリング誤り率

動画	誤クラスリングフレーム数	クラスタリング誤り率[%]
The Beatles の全 4 動画	46	0.5
宇多田ヒカルの全 4 動画	149	2.5

この結果から，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，本研究のフレームの時間的連続性と顔類似度を用いた枠組みによって音楽連動動画における顔認証の精度が向上していることがわかる．

表 4　歌声区間における映像中のボーカルの登場率

動画名	総フレーム数	歌声区間のフレーム数	歌声区間でのボーカル登場率[%]	歌声区間以外でのボーカル登場率[%]
Let it be / The Beatles	7097	5348	60.1	12.5
Hey Jude / The Beatles	7196	6028	75.1	6.5
Get Back / The Beatles	5459	3025	37.2	15.3
Two of us / The Beatles	6218	4502	64.1	24.1
Wait & See / 宇多田ヒカル	6655	6052	61.8	3.8
For You / 宇多田ヒカル	6034	5305	93.0	11.7
Can You Keep A Secret? / 宇多田ヒカル	5931	5928	70.9	0.0
Final Distance / 宇多田ヒカル	4924	4194	66.4	2.5

5. 音楽連動動画における歌声とボーカルの顔の関係

音楽連動動画における，時系列情報を持ったアノテーションの種類には様々なものが考えられる. 例えば, A メロやサビなどと呼ばれるような楽曲の繰り返し構造や，映像のシーンの繰り返し構造などである．これらの音楽の時系列情報と映像の時系列情報の間に関係性を見出すことができれば，それを利用したアノテーションの精度向上も期待できる．

ここで，映像中に登場するアーティストの顔に影響することが考えられる要素として，音楽の歌声区間と登場するアーティストとの関係について考える．一般的に歌声区間では，ボーカルの顔が映っている可能性が高いことが予想できる．そこで予備実験として，実際に実験で使用した音楽連動動画について，手作業で歌声区間をラベリングして，その歌声区間と，ボーカルの顔が映っているフレームとの関係を調べた．結果を表 4 に示す．ここで，ボーカルの登場率は以下の式(5),(6)のように算出した．

$$\text{歌声区間でのボーカル登場率} = \frac{\text{ボーカル登場フレーム数}}{\text{歌声区間のフレーム数}} \quad (5)$$

$$\text{歌声区間以外でのボーカル登場率} = \frac{\text{ボーカル登場フレーム数}}{\text{歌声以外の区間のフレーム数}} \quad (6)$$

この分析結果より，歌声区間ではボーカルの顔が映像に登場していることが多いと言える．The Beatles は，ボーカルが歌だけでなく楽器も担当しているために，ボーカルパート以外でも楽器奏者として登場する場合も多く存在している．それを踏まえて

Vol.2012-MUS-94 No.24
Vol.2012-SLP-90 No.24
2012/2/4

も，歌声区間ではボーカルの顔が登場していることが多いと言える．

この結果から，歌声区間を含む顔時間連続体には，正解顔画像群の中に存在するすべてのボーカルの画像を正解としやすくするような重み付け処理を加える方法の有効性が示唆される．ただし，現状では作品によってボーカル登場率に違いがあるため，そういった作品毎の演出などの違いを考慮する必要がある．今回の分析では，歌声区間の手動ラベリングによりデータを分析したが，楽曲の歌声区間検出 [15](VAD: Voice Activity Detection)を用いることで本アノテーション付加手法を拡張することができる．

6. おわりに

本研究では，音楽動画コンテンツ（音楽連動動画）に対して，マルチフレーム認識の枠組みに基づく人物の顔に特化した認証手法を提案し，アーティスト名とその登場シーンの同定でその有効性を確かめた．音楽連動動画の利用においてアーティストは重要な情報であり，さらなる同定精度向上のために，本同定手法を拡張していく予定である.例えば,映像の内容理解の対象を人物の体全体や周辺の物体にまで広げたり，顔時間連続体を人物が後ろを向くまでトラッキングすることで，従来は扱うことが困難だった後ろ向きの人物の顔認証を実現したりすることを目指したい．

現在の実装では，顔時間連続体の正解顔画像を含む正解顔画像群を手作業で用意した後に顔認証を行っているが，本来はインターネット上の膨大な顔画像群を検索して活用することを想定している．今後は，そうしたインターネット上の顔画像とそのラベルに基づいてアーティスト名を同定する予定である．また，既にインターネット上の音楽連動動画に付与されているソーシャルアノテーションから映像中のアーティスト候補を得て，その登場確率を高める等，さらなる精度向上に取り組みたい．

第 5 章でも歌声区間とボーカルの顔との関連から，歌声区間を自動推定して精度向上に利用できる可能性を考察したが，歌声の声質を推定 [16]することで，ボーカルの候補を絞ったり，ギターソロの区間を推定してギタリストの候補を絞ったりできる可能性がある．我々は，このような音楽内容と映像との関係性を活用する研究にも既に着手しており，人物の顔以外も含め，幅広く本アプローチを発展させていきたいと考えている．

謝辞 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．

参考文献

1) Google Videos：http://www.google.com/videohp
2) ニワンゴ：ニコニコ動画，http://www.nicovideo.jp/
3) 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 Vol.52，No.12，pp.3471-3482 (2011).
4) TRECVID, http://www-nlpir.nist.gov/projects/trecvid/
5) 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，情処研報，2011-CVIM-177，pp.1-8 (2011).
6) 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™」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 (2005).
7) 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).
8) 木下航一，小西嘉典，勞世竑，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，No.4，pp.721-729 (2011).
9) Irie, A., Takagiwa, M., Moriyama, K., and Yamashita, T. : Improvements to Facial Contour Detection by Hierarchical Fitting and Regression，Asian Conference on Pattern Recognition (ACPR) (2011).
10) Blanz, V., Mehl, A., Vetter, T., and Seidel, H.：A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data，Proc. Symp. on 3D Data Processing，Visualization and Transmission，pp.293-300 (2004).
11) Dalal, N. and Triggs, B.：Histograms of Oriented gradients for human detection，Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，pp.886-893 (2005).
12) 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(361)，pp.117-122 (2011).
13) Lowe, D. G.：Object Recognition from Local Scale-Invariant Features，Proc. IEEE International Conference on Computer vision (ICCV)，pp.1150-1157 (1999).
14) Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. A.：Who's in the Picture，Proc. Neural Information Processing Systems (NIPS) ，pp.137-144 (2004).
15) 藤原弘将，北原鉄朗，後藤真孝，駒谷和範，尾形哲也，奥乃博：伴奏音抑制と高信頼度フレーム選択に基づく楽曲の歌手名同定手法，情処論 Vol.47，No.6，pp.1831-1843 (2006).
16) Fujihara, H. and Goto, M.：A music information retrieval system based on singing voice timbre，Proc. ISMIR 2007，pp. 467–470 (2007).